On Modeling Spatial Time-to-Event Data with Missing Censoring Type

Diane Lu

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy under the Executive Committee of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2024
Abstract
On Modeling Spatial Time-to-Event Data with Missing Censoring Type
Diane Lu

Time-to-event data, a common occurrence in medical research, is also pertinent in the ecological context, exemplified by leaf desiccation studies using innovative optical vulnerability techniques. Such data can unveil valuable insights into the influence of various factors on the event of interest. Leveraging both spatial and temporal information, spatial survival modeling can unravel the intricate spatiotemporal dynamics governing event occurrences. Existing spatial survival models often assume the availability of the censoring type for censored cases. Various approaches have been employed to address scenarios where a "subset" of cases lacks a known "censoring indicator" (i.e., whether they are right-censored or uncensored). This uncertainty in the subset pertains to missing information regarding the censoring status. However, our study specifically centers on situations where the missing information extends to "all" censored cases, rendering them devoid of a known censoring "type" indicator (i.e., whether they are right-censored or left-censored).

The genesis of this challenge emerged from leaf hydraulic data, specifically embolism data, where the observation of embolism events is limited to instances when leaf veins transition from water-filled to air-filled during the observation period. Although it is known that all veins eventually embolize when the entire plant dries up, the critical information of whether a censored leaf vein embolized before or after the observation period is absent. In other words, the censoring
type indicator is missing.

To address this challenge, we developed a Gibbs sampler for a Bayesian spatial survival model, aiming to recover the missing censoring type indicator. This model incorporates the essential embolism formation mechanism theory, accounting for dynamic patterns observed in the embolism data. The model assumes spatial smoothness between connected leaf veins and incorporates vein thickness information. Our Gibbs sampler effectively infers the missing censoring type indicator, as demonstrated on both simulated and real-world embolism data. In applying our model to real data, we not only confirm patterns aligning with existing phytological literature but also unveil novel insights previously unexplored due to limitations in available statistical tools. Additionally, our results suggest the potential for building hierarchical models with species-level parameters focusing solely on the temporal component. Overall, our study illustrates that the proposed Gibbs sampler for the spatial survival model successfully addresses the challenge of missing censoring type indicators, offering valuable insights into the underlying spatiotemporal dynamics.
# Table of Contents

Acknowledgments ........................................................................................................ viii

Chapter 1: Introduction .................................................................................................. 1
  1.1 Motivating applications ......................................................................................... 1
  1.2 Challenges with such data ..................................................................................... 2
  1.3 Contributions ........................................................................................................ 2

Chapter 2: Background ................................................................................................ 4
  2.1 Spatial survival models ......................................................................................... 4
  2.2 Methods dealing with missing censoring .............................................................. 5
  2.3 Main Application .................................................................................................. 6

Chapter 3: Gibbs algorithm .......................................................................................... 7
  3.1 Notation ................................................................................................................ 7
    3.1.1 Complete Data .................................................................................................. 7
    3.1.2 Doubly Censored Data ..................................................................................... 7
    3.1.3 Observed Data ................................................................................................ 8
  3.2 Gibbs Sampler for the Missing Censoring Type Indicator Problem .................... 10
    3.2.1 Model Set-Up .................................................................................................. 10
    3.2.2 Gibbs Sampler ................................................................................................. 14
3.2.3 Posterior Inference ........................................... 18

3.3 Hierarchical Model .................................................. 20
  3.3.1 Model Set-Up .................................................. 20
  3.3.2 Gibbs Sampler .................................................. 22

Chapter 4: Application .................................................. 26
  4.1 Machine Learning Workflow for Embolism Event Detection ................. 26
    4.1.1 Noises ...................................................... 27
    4.1.2 Machine Learning Workflow ................................ 29
    4.1.3 Stem Data .................................................. 32
    4.1.4 Implementation Details ...................................... 33
    4.1.5 Performance Evaluation ...................................... 34
  4.2 Exploratory Data Analysis ............................................. 35
    4.2.1 Survival Model for Embolism Data: A Rationale ..................... 35
    4.2.2 Leaf Data .................................................. 37
    4.2.3 Limitation of Only Considering Right-censoring ....................... 43
  4.3 Application of the Gibbs ........................................... 45
    4.3.1 Simulation: Parameter and Censorship Type Indicator Recovery ........ 45
    4.3.2 Simulation: Robustness Analysis ................................ 48
    4.3.3 Simulation: Empirical Bias Evaluation ................................ 49
    4.3.4 Real-world Data ............................................. 55

Conclusion .............................................................. 64
References ................................................................. 68

Appendix A: Within-species Correlation ............................. 72

Appendix B: Shapiro-Wilk Normality Test .......................... 76

Appendix C: Prior and Posterior Distributions ..................... 77

Appendix D: Choice of Survival Model for Modeling Embolism Events ....................... 79
List of Figures

3.1 The relationship between the complete list of event times and that within the observation period. The event times in the first line refers to the sorted event times from the complete data, where as the ones in the second line refers to the sorted "observed" event times from the observed data. .................................................. 9

3.2 The graphical representation of our model. .................................................... 13

3.3 Subsets of neighbors introduced for updating the spatial frailties. Solid subjects’ spatial frailties are being updated given the spatial frailties of those marked by "X". 17

3.4 The graphical representation of the hierarchical model. ............................... 22

4.1 Different sources of noise on optical vulnerability image data. ...................... 28

4.2 The machine learning workflow for detecting embolism events using optical vulnerability stem image data. ......................................................... 29

4.3 Different polynomial transformation on measured water potential. .............. 39

4.4 An example for vein segment processing. (a) Unprocessed image of a *Ocotea leucoxylon* leaf. (b) Binarized image derived from the first image using phenoVein. (c) Locations of the branching points and ending points are highlighted by yellow and red markers respectively through phenoVein. (d) Visualization of embolism events during desiccation in the observation period (i.e, increases in water potential, wp). 41

4.5 (a) A grayscale image example of a vein segment. (b) The binarized image of the vein segment. (c) The distance transform of the vein segment visualizes how far each vein pixel is from the background pixel. The pixels with higher values (yellow pixels) form the medial axis. The maximum Euclidean distance of the pixels on the medial axis to the nearest background pixel is 5.83, and hence the thickness for this vein segment is estimated as $2 \times 5.83 = 11.66$. 42
4.6 Example of left-censoring in a Casearia arborea leaf. (a) Observed embolism events. Dashed circle highlights where left-censoring is suspected to occur. The color red indicates that the vein segment embolizes earlier, whereas purple signifies a later embolism occurrence. (b) Posterior mean of log odds ratio if we only assume right-censoring. Notice that the circled midvein is predicted to be right-censored.

4.7 The graphical representation of our model under the empirical bias framework.

4.8 The average time per iteration (s) for our Gibbs sampler was evaluated under the empirical bias study across different numbers of events per vein segment (N). The dashed line represents the quadratic regression line, with the p-value for the quadratic term being 0.0224, indicating statistical significance. This suggests that the average time increases quadratically with the increase of N.

4.9 The posterior mean of \( \tau^2 \) obtained from our Gibbs sampler was evaluated across different numbers of events per vein segment (N) in an empirical bias study. The solid line represents the true value of the spatial variance parameter. Notably, our estimator’s bias decreases as N increases.

4.10 Gibbs sampler result on the Casearia arborea leaf showing the relative posterior mean of log odds ratio. The color gradient from red to purple indicates the progression of embolism, where red represents earlier stages and purple represents later stages. Additionally, red indicates left-censoring, while the turquoise to purple spectrum indicates right-censoring.

4.11 The box plots display parameter values grouped by species. (Upper Row and Bottom Left) Species are ordered by Weibull mean, enabling an exploration of the correlation between different parameters and the fundamental temporal component. (Bottom Right) Species are ordered by mean vein thickness coefficient, facilitating an inspection of the correlation between the spatial variance parameter and vein thickness coefficient.

4.12 Relationship between mean \( P_{50} \) (MPa) from [35] and posterior mean from the Gibbs sampler: (Upper Left) Vein thickness coefficient. (Upper Right) The inverse of the spatial variance parameter. (Bottom Left) The fundamental temporal component characterized by the mean of Weibull family in log scale. The dashed line shows the linear regression line.

4.13 Relationship between vein thickness regression coefficient from our spatial survival model and (a) Average vein density and (b) Degree of connectivity. The dashed line shows the linear regression results.

4.14 The unifying machine learning workflow for studying embolism events using optical vulnerability image data.
A.1 Highest posterior density regions for $\beta$ and Weibull mean in log scale. . . . . . . . 73
A.2 Highest posterior density regions for $\tau^2$ and Weibull mean in log scale. . . . . . . . 74
A.3 Highest posterior density regions for $\tau^2$ and $\beta$. . . . . . . . . . . . . . . . . . . . . 75

C.1 The prior distribution, the posterior mean, and the posterior standard deviation for each parameter. The black curve signifies the prior distribution, while the colored circles denote the posterior means. The colored error bars represent one standard error around the posterior means. . . . . . . . . . . . . . . . . . . . . . 78

D.1 The box plots display the relative improvement of the proportional hazards model compared to the proportional odds model, grouped by species. . . . . . . . . . . . 80
List of Tables

4.1 Data Summary. ................................................................. 32

4.2 Performance of the proposed machine learning workflow using different ways of foreground background segmentation. ......................................................... 34

4.3 List of species used in this study, code used in figures, their families, and number of leaves measured per species. ................................................................. 38

4.4 Results for predicting the missing censoring type indicator. In the confusion matrix, left-censored, right-censored, and uncensored are abbreviated as "l", "r", and "u" respectively. The overall accuracy encompasses both censored and uncensored cases, whereas censored accuracy specifically considers only censored cases. ... 46

4.5 The hyperparameter values for the simulated dataset. ......................................................... 47

4.6 The simulated dataset’s true parameter values. The posterior mean and 90% interval obtained through our proposed Gibbs sampler and through the model only assumes right-censoring. ......................................................... 47

4.7 Results for predicting the missing censoring type indicator on the data without any left-censoring instances. In the confusion matrix, left-censored, right-censored, and uncensored are abbreviated as "l", "r", and "u" respectively. The overall accuracy encompasses both censored and uncensored cases, whereas censored accuracy specifically considers only censored cases. ......................................................... 50

4.8 The true parameter values of the simulated dataset, which is devoid of left-censored instances. The posterior mean and 90% interval obtained through our proposed Gibbs sampler and through the model only assumes right-censoring. ............... 50

4.9 Kruskal-Wallis test on examining the differences on each parameter among various species. ................................................................. 62

B.1 Shapiro-Wilk normality test on posterior mean of parameters and vein features. ... 76
Acknowledgements

I am immensely grateful to my advisor, Prof. Tian Zheng, whose unwavering guidance, steadfast support, and insightful feedback have been indispensable throughout the entire research process. Prof. Zheng not only generously shared her expertise in the subject matter but also provided invaluable advice on maintaining a healthy work-life balance and offered meaningful suggestions on effective mentorship. Her multifaceted support has played a pivotal role in shaping the direction of this thesis.

Sincere appreciation goes to the members of my defense committee—Prof. Maria Uriarte, Prof. Zhiliang Ying, Prof. Mengling Liu, and Prof. Yuqi Gu—for dedicating their time, expertise, and thoughtful contributions during the defense of my thesis. I extend special thanks to Prof. Maria Uriarte for her invaluable insights and expertise in phytological research, which significantly enriched the examination and discussion of my work.

Heartfelt gratitude is extended to my collaborators, Fangyi Chen, Chris M. Smith-Martin, Dingyi Fang, and Prof. Robert Muscarella, for engaging in fruitful discussions on our research endeavors. I also express my appreciation to the entire faculty and staff in the Department of Statistics. In addition, a sincere thank you to all my friends who have accompanied me on this challenging yet rewarding Ph.D. journey.

My deepest gratitude goes to my family. Their unwavering support, love, and encouragement provided the hope and strength I needed, especially during moments when I doubted myself. Overcoming obstacles became possible because of their constant presence in my life.
Chapter 1: Introduction

1.1 Motivating applications

In the pursuit of scientific inquiry, we often encounter complex scenarios that necessitate the investigation of events with indeterminate temporal attributes. Two distinct applications serve to illustrate this pervasive challenge, where uncertainties in the timing of critical events present formidable analytical hurdles.

**Embolism Example:** Let us consider an ecological scenario, one that endeavors to unravel the enigmatic process of leaf dehydration in extreme climatic conditions, such as drought. In this context, we possess the capacity to ascertain whether the leaf veins have transitioned from a water-filled to an air-filled state. However, the precise state, whether it remains water-filled or has transformed into air-filled, eludes our observations. Crucially, it is known that once a vein succumbs to air blockage, it perpetually maintains this condition, thereby impairing its ability to transport water—akin to a vital conduit having succumbed to an irreversible impairment. Over time, the cumulative consequence of air-blocked veins portends the eventual demise of the entire plant, as all veins ultimately transit to this "embolism" (air-blocked) state.

**Medical Example:** In an alternative medical scenario, our investigation centers on the study of COVID-19, a coronavirus disease. Initially, Polymerase Chain Reaction (PCR) tests were developed to diagnose active coronavirus infections. However, due to their high cost, PCR tests were only conducted during a specific observation period. Within this observation period, some individuals did not exhibit noticeable symptoms of illness (excluding the possibility of asymptomatic cases for the purpose of focusing on the missing censoring problem). Interestingly, couple of months after the observation period ended, a breakthrough occurred when antibody tests were developed, allowing for a retrospective determination of whether an individual had contracted the
disease in the past. Surprisingly, it is revealed that every member within the study had indeed been infected.

1.2 Challenges with such data

In both of these compelling scenarios, the underlying events transpire with a temporally transient nature akin to abrupt spikes, leaving us deprived of the crucial insight required to ascertain their current status. In the case of the leaf embolism example, for those veins that did not undergo air-blockage during the defined observation period, a fundamental ambiguity prevails—whether they transformed into air-filled conduits prior to the commencement of the observation window or after its termination. Likewise, in the context of the COVID-19 investigation, the individuals who were not infected during the observational period pose an enigma; it remains indeterminate whether their affliction occurred antecedent to or subsequent to the observation period.

This prevailing ambiguity presents a formidable challenge in the estimation of survival probabilities. The crux of the matter lies in the inherent assumption of standard survival analysis: the a priori knowledge of all unobserved events. In essence, the pivotal question that emerges from these perplexing scenarios is how one may model the survival probability in the presence of missing priori knowledge on the unobserved events.

1.3 Contributions

In response to the aforementioned challenges, we present a novel approach grounded on Bayesian modelling. Specifically, we employ Gibbs sampling, a well-established technique renowned for its efficacy in managing datasets with a high dimensionality, which is often encountered in scenarios featuring an abundance of variables. Our developed Gibbs sampler affords us the capability to make precise inferences regarding data with missing information on the unobserved events, thus circumventing the inherent uncertainties arising from data gaps.

We substantiate the effectiveness of our proposed methodology through a comprehensive evaluation, employing both simulated data and real-world data. This rigorous assessment showcases
the reliability and applicability of our solution, ultimately contributing to the advancement of our understanding in scenarios characterized by temporal ambiguities and missing data.
Chapter 2: Background

2.1 Spatial survival models

Spatial survival analysis is a specialized branch of statistical analysis that combines principles of survival analysis with spatial statistics to study the time-to-event data of individuals or entities in the context of spatial or geographic information. In traditional survival analysis, the primary focus is on understanding the time until an event of interest occurs, such as death, failure, or disease progression. In spatial survival analysis, this temporal aspect is enriched by considering the spatial distribution of events and how it influences the survival experience.

This field is particularly valuable in numerous domains, including epidemiology, ecology, and public health [1, 2, 3], where the geographical location of individuals or events can play a critical role in influencing the time to an event. For example, in epidemiology, spatial survival analysis can help researchers understand how the geographic distribution of individuals affects the spread of diseases and the timing of outbreaks [4, 5]. In ecology, it can be used to analyze the survival times of animals or plants in spatially varying environments[6, 7, 8, 9].

In the realm of survival analysis, the Cox proportional hazard model [10] stands as the most widely employed approach. However, there are specific scenarios where the Cox model may not be the most suitable choice. The Cox model operates under the assumption that covariates exert a consistent, fixed multiplicative influence on the hazard rate.

Acknowledging the limitations of this assumption, it may be more realistic to consider situations where the influence of covariates on the hazard rate diminishes over time [11]. In response to this, the proportional odds model was introduced [12] as an alternative within the domain of survival analysis. In contrast to the Cox model, which presumes that the ratio of hazards associated with different covariate values remains constant over time, the proportional odds model relaxes
this assumption. It allows the hazard ratio to vary over time while maintaining the odds ratio as a fixed parameter throughout time.

Hence, our study focuses on the use of the proportional odds model with spatial covariates. It’s worth noting that spatial proportional odds models have demonstrated utility in various domains, including cancer research [13] and the study of pedestrian crash injuries [14].

2.2 Methods dealing with missing censoring

Correct censoring type is of paramount importance in survival analysis because the choice of censoring mechanism directly influences the accuracy of estimations related to the probability of an event occurring. Inaccurate or improper censoring can lead to biased estimates and potentially distort the understanding of the event occurrence over time. Two notable references, [15] and [16], have highlighted the critical impact of censoring on the accuracy of survival analysis, underlining the significance of selecting the appropriate censoring type to ensure that the estimated probabilities are reliable and trustworthy.

The issue presented below refers to a different problem compared to our own, but it shares the characteristic of being a missing data problem within the domain of survival analysis. In the scenario denoted as the Missing Censoring Indicator (MCI) model, a subset of the data exhibits uncertainty regarding whether the observations are right-censored or uncensored. However, for the remaining data, the censoring indicators are available and certain. Several diverse strategies have been proposed to address this challenge. These include the utilization of kernel density estimation techniques, as demonstrated in [17], imputation methods, as explored in [18], and Bayesian approaches, exemplified by [19]. These methodologies aim to handle the uncertainty associated with censoring indicators and contribute to the robust analysis of survival data in the presence of missing information.

However, it is worth noting that, to date, existing scholarly works have not explored scenarios in which the "censoring type indicator" is absent for the entire data set. Specifically, in such instances, the determination of whether the censoring is of the left-censoring or right-censoring...
variety remains unknown. Our knowledge is restricted to the occurrence of the event of interest within the observation period. Nevertheless, for the unobserved cases, the categorization into left-censoring or right-censoring status remains unresolved. This presents a hitherto unaddressed challenge within the realm of survival analysis.

2.3 Main Application

Our introduction to this unique challenge arose within the context of leaf hydraulic data analysis. The challenge stems from the fact that we can only observe embolism events when there’s a change in reflectance, indicating the transition of veins from water-filled to air-filled. This observational approach leaves us unable to discern the current state of each vein, and as a result, we lack the information to determine whether the unobserved veins fall into the category of left-censoring or right-censoring. However, by combining insights from existing phytological literature and incorporating the assumption of spatial frailty smoothness, we endeavor to infer the missing censoring type indicators.

Our approach is guided by several observations within our dataset. Notably, we have observed that most of the censored veins are characterized by their thinness, with only a small proportion of these being midveins (i.e., the thickest veins within a leaf). Previous phytological studies have highlighted the tendency for thicker veins to experience embolism at an earlier stage, as documented in works such as [20, 21, 22, 23, 24]. Drawing from this knowledge, we surmised that a limited degree of left-censoring may be evident among censored midveins, while a substantial predominance of right-censoring may be present among the censored thin veins.
Chapter 3: Gibbs algorithm

3.1 Notation

3.1.1 Complete Data

In our framework, we begin by establishing $m$ as the total number of observed subjects. Each of these subjects is associated with an event time denoted as $t_i$ and is characterized by a covariate $v_i$. Furthermore, we consider the spatial relationships between subjects and represent them using an adjacency matrix denoted as $E = \{e_{ij}\}$. Specifically, in the adjacency matrix, $e_{ij} = 1$ if the $i^{th}$ and $j^{th}$ subjects are considered neighbors, and $e_{ij} = 0$ otherwise. This adjacency matrix, denoted as $E$, serves to capture the spatial dependence between subjects in the study.

The complete data, reflecting all the available information, can be concisely summarized as $X^0 := \{(v_i, t_i)\}_{i=1}^m \cup E$. This comprehensive representation includes the covariate values, event times, and spatial relationships.

3.1.2 Doubly Censored Data

However, we might not be able to observe all the events in practice. So, let’s consider a scenario where we observe a total of $n$ subjects, each of whom experiences an event within the specified observation period $(T_1, T_2)$. It is crucial to acknowledge that $n$ can be either less than or equal to $m$. Doubly censored data [25] encompasses events that are categorized as either left-censored, right-censored, or uncensored. In this context, we have complete knowledge regarding the type of censoring for each event.

The recorded temporal information for doubly censored data is denoted as $t_i^\ast$. It signifies the event time for the $i^{th}$ subject when an observed event takes place within the observation period. If no event is observed for the $i^{th}$ subject, $t_i^\ast$ represents the censoring time. Specifically, for right-
censoring, the censoring time is $T_2$, and for left-censoring, it is $T_1$. The relationship between these variables can be expressed as follows:

$$\forall i \in [m], \ t^*_i = \begin{cases} 
  t_i & \text{if } t_i \in (T_1, T_2) \\
  T_2 & \text{if } t_i \geq T_2 \\
  T_1 & \text{if } t_i \leq T_1 
\end{cases}$$

$$= \max(\min(t_i, T_2), T_1)$$

Here, $t^*_i$ is defined for each subject, and it is determined based on the subject’s event time $t_i$, accounting for whether it falls within the observation period. The doubly censored data is denoted as $X^* := \{(v_i, t^*_i)\}_{i=1}^m \cup E$.

3.1.3 Observed Data

In our observed data $X := \{(v_i, t^{|o}_i)\}_{i=1}^m \cup E$, the observed event time for subjects who are not censored corresponds to the event time $t_i$. For subjects where no event was observed (i.e., censored), $t^{|o}_i$ is absent, as illustrated below:

$$\forall i \in [m], t^{|o}_i = \begin{cases} 
  t_i & \text{if } t_i \in (T_1, T_2) \\
  NA & \text{else} 
\end{cases}$$

Here, the "NA" value indicates missing data for subjects who did not experience an event during the observation period.

Denote the sorted (in ascending order) event times as $\{t(1), ..., t(m)\}$ and the sorted observed event times as $\{t^{|o}_1, ..., t^{|o}_n\}$. A graphical representation of the relationship between the complete list of event times and those observed within the observation period can be observed in Figure 3.1. In situations where all events occurring within the observation period $(T_1, T_2)$ are observable, the
Figure 3.1: The relationship between the complete list of event times and that within the observation period. The event times in the first line refers to the sorted event times from the complete data, whereas the ones in the second line refers to the sorted "observed" event times from the observed data.

observed event time for the $i^{th}$ subject can be expressed as follows:

$$\forall i \in [n], \quad t_{(o)}^{(i)} = t_{(N_{T_1}+i)}I(t_{(N_{T_1}+i)} \in (T_1, T_2))$$

Here, $N_{T_1}$ represents the count of events occurring before $T_1$, and $I(t_{(N_{T_1}+i)} \in (T_1, T_2))$ is an indicator function ensuring that the observed event time falls within the specified observation interval.

In our observed data, we maintain a structure similar to that of doubly censored data, with the notable difference that we are missing the information on censoring type. Specifically, we do not know whether the censored subjects are left-censored or right-censored. It’s important to note that, when the censoring type (i.e., left- or right-censoring) is known, we can transform the observed data into the format of standard doubly censored data.
3.2 Gibbs Sampler for the Missing Censoring Type Indicator Problem

3.2.1 Model Set-Up

The survival and density functions for proportional odds models (PO models [12]) of \(i^{th}\) subject are

\[
S_{\beta,\gamma,r_i}(t) = \frac{e^{-(r_i + \beta v_i)} S_\gamma(t)}{1 + (e^{-(r_i + \beta v_i)}) - 1 S_\gamma(t)} \quad (3.1a)
\]

\[
f_{\beta,\gamma,r_i}(t) = \frac{e^{-(r_i + \beta v_i)} f_\gamma(t)}{[1 + (e^{-(r_i + \beta v_i)}) - 1 S_\gamma(t)]^2} \quad (3.1b)
\]

Here, the individual components are defined as follows:

1. Covariate Coefficient \(\beta\): This parameter, denoted by \(\beta\), quantifies the impact of the covariate on the event occurrence. A positive value of \(\beta\) signifies that as the covariate value increases, the likelihood of the event also increases.

2. Unobserved Spatial Frailty \(r_i\): The spatial frailty, denoted as \(r_i\), is associated with each subject and reflects their individual susceptibility to experiencing the event. A higher value of the spatial frailty suggests a greater propensity for the event to occur earlier. The spatial component of \(r_i\) is modeled using the conditional autoregressive (CAR) prior, which incorporates information from the adjacency matrix and assumes spatial smoothness. The CAR prior is characterized by the spatial variance parameter \((\tau^2)\), governing the variability of spatial frailty.

3. Fundamental Temporal Component \(S_\gamma(t)\): This function encodes the survival probabilities corresponding to \(r_i + \beta v_i = 0\) (e.g., when there’s no spatial nor covariate effect). To model the survival function, we adopt the Weibull family, which is characterized by the scale parameter \(\frac{1}{e^{y_1}}\) and the shape parameter \(e^{y_2}\). Hence, \(\gamma := (y_1, y_2)\) Our choice of the Weibull family is informed by its widespread application in leaf hydraulic analysis[26, 27]. The associated probability density function (PDF) and survival function for the Weibull family are mathe-
matically defined as \( f_\gamma(t) = e^{\gamma_1 + \gamma_2 (e^{\gamma_1 t})} e^{\gamma_2 - 1} e^{-(e^{\gamma_1 t})} e^{\gamma_2} \) and \( S_\gamma(t) = e^{-(e^{\gamma_1 t})} e^{\gamma_2} \), respectively.

In a broader context, the choice of a parametric family for the fundamental temporal component can vary based on the specific requirements and objectives of the application at hand. Different parametric families offer various modeling capabilities and assumptions, and selecting the most suitable one depends on the characteristics of the data and the goals of the analysis. Researchers or practitioners can tailor their choice to align with the nuances of their particular research domain and the phenomena they intend to investigate. The flexibility to select an appropriate parametric family allows for the adaptation of survival models to diverse applications, ensuring that the modeling approach is well-suited to the context and the underlying data.

The likelihood depends on the survival and density functions:

\[
P(t_i^{[o]} | \beta, \gamma, r_i) = \begin{cases} 
  f_{\beta, \gamma, r_i}(t_i^{[o]}) & \text{if } t_i^{[o]} \neq NA \\
  P(t_i \notin (T_1, T_2) | \beta, \gamma, r_i) & \text{else}
\end{cases}
\]

\[
= \begin{cases} 
  f_{\beta, \gamma, r_i}(t_i^{[o]}) & \text{if } t_i^{[o]} \neq NA \\
  1 - S_{\beta, \gamma, r_i}(T_1) + S_{\beta, \gamma, r_i}(T_2) & \text{else}
\end{cases}
\]

In instances where the observed event time \( t_i^{[o]} \) is marked as NA (i.e., missing data), it is imperative to account for both scenarios of left-censoring and right-censoring. This consideration is necessary due to the absence of the censoring type indicator, requiring an exploration of both possibilities.
The prior distributions are specified as below:

\[ \beta \sim N(\beta_0, B_0) \]
\[ \gamma_1 \sim N(\gamma_{0,1}, V_{0,1}) \]
\[ \gamma_2 \sim N(\gamma_{0,2}, V_{0,2}) \]
\[ \tau^{-2} \sim \Gamma(a_\tau, b_\tau) \]

\[ r_{-q} := (r_1, \ldots, r_{q-1}, r_{q+1}, \ldots, r_m) \top | \tau^2 \sim CAR_{m-1}(\tau^2) = N_{m-1}(0, \tau^2(D_{-q} - \alpha_r E_{-q})^{-1}) \]
\[ r_q = -\sum_{i\neq q} (r_i) \]

Here, the gamma distribution is denoted as \( \Gamma(a, b) \), where \( a \) represents the shape parameter, and \( b \) is the rate parameter (with a mean equal to \( \frac{a}{b} \)). The spatial frailty vector, denoted as "r", is subject to a CAR prior with zero-mean constraint. In this distribution, \( m - 1 \) represents the dimension of \( r_{-q} \), while the covariance matrix’s structure relies on several parameters. These parameters include the spatial variance parameter \( \tau^2 \), which quantifies the variability in the spatial frailty; the degree matrix \( (D) \), a diagonal matrix where each diagonal element reflects the number of neighbors for a given subject; the spatial correlation parameter \( (\alpha_r) \), and the adjacency matrix \( (E) \), which encodes the spatial relationships among subjects. The parameter \( (\alpha_r \in (0, 1)) \) determines the spatial correlation, a value of 0 means subjects are spatially independent, while 1 means complete spatial correlation.

Under the usual CAR prior (i.e., \( r_1, \ldots, r_m | \tau^2 \sim CAR_m(\tau^2) = N_m(0, \tau^2(D - \alpha_r E)^{-1}) \)), the individual spatial frailty term \( r_i \) and the covariate coefficient \( \beta \) are not uniquely identifiable. Only the log odds ratio, represented as \( r_i + \beta v_i \), is identifiable. This means that if we were to consider a scenario where \( \beta' \) is greater than \( \beta \) and we define a new spatial frailty term \( r'_i \) as \( r'_i = r_i + \beta v_i - \beta' v_i \), we would end up with the same log odds ratio. Consequently, this results in identical survival functions, survival densities, and likelihoods. This uniformity arises because the survival functions and densities are contingent on \( r_i \) and \( \beta \) solely through the log odds ratio, making them invariant to changes in the individual values of \( r_i \) and \( \beta \).
To address the identifiability issue, we introduced a zero-mean constraint on the spatial frailty prior. This involved replacing the usual CAR prior distribution by

$$r_{-q} | \tau^2 \sim N_{m-1}(0, \tau^2 (D_{(-q)} - \alpha r E_{(-q)})^{-1})$$

$$r_q = - \sum_{i \neq q} (r_i)$$

Here, $q$ represents the index of an "uncensored" subject. It’s important to note that if $q$ denotes a "right-censored" subject, it can lead to a similar issue where $\beta$ becomes overestimated, $r_{-q}$ becomes underestimated, $r_q$ becomes overestimated (to satisfy the zero-mean constraint), and $\tau^2$ becomes overestimated. The overestimation of $r_q$ should not pose a problem because the $q^{th}$ subject is right-censored. Thus, it’s acceptable for the log odds ratio to be unreasonably large. However, when $q$ corresponds to an "uncensored" subject, it’s crucial to avoid overestimating $r_q$ to an excessively high value. This is because an excessively large $r_q$ could lead to an extremely small value for $P(t_{i(q)}^{[o]} | \beta, \gamma, r_q)$, which is problematic.

The graphical model representation is illustrated in Figure 3.2. The number of parameters ($\beta, \gamma, \tau^2$) do not vary with sample size, while the number of missing latent spatial frailties ($r_i$) and the observed event times ($t_{i(q)}^{[o]}$) increase as the sample size ($m$) increases. The dependencies are also evident through the graphical model. The spatial frailties ($r_i$) are controlled only by the spatial variance parameter $\tau^2$, and the observed event time ($t_{i(q)}^{[o]}$) depends on spatial frailty ($r_i$), the
fundamental temporal component governed by $\gamma$, and the covariate coefficient $\beta$ through Equation 3.1.

3.2.2 Gibbs Sampler

Joint distribution of hidden and observed variables:

$$P(\beta, \gamma, \tau^2, r, t^{[o]}) = \prod_{i=1}^{m} P(t_i^{[o]}|\beta, \gamma, r_i)P(r_{-q}|\tau^2)P(\tau^2)P(\gamma)P(\beta)$$

In the Gibbs sampler, each sample is drawn from the full conditional distribution while conditioning on the latest values of all other variables. For the $(s+1)^{th}$ iteration of the Gibbs sampler,

$$\beta^{(s+1)} \sim P(\beta^{(s+1)}|\gamma^{(s)}, \tau^{2(s)}, r^{(s)}, t^{[o]})$$

$$\gamma_1^{(s+1)} \sim P(\gamma_1^{(s+1)}|\beta^{(s+1)}, \gamma_2^{(s)}, \tau^{2(s)}, r^{(s)}, t^{[o]})$$

$$\gamma_2^{(s+1)} \sim P(\gamma_2^{(s+1)}|\beta^{(s+1)}, \gamma_1^{(s+1)}, \tau^{2(s)}, r^{(s)}, t^{[o]})$$

$$\tau^{2(s+1)} \sim P(\tau^{2(s+1)}|\beta^{(s+1)}, \gamma^{(s+1)}, r^{(s)}, t^{[o]})$$

$$r_i^{(s+1)} \sim P(r_i^{(s+1)}|\beta^{(s+1)}, \gamma^{(s+1)}, \tau^{2(s+1)}, r_1^{(s+1)}, \ldots, r_i^{(s+1)}, \ldots, r_i^{(s+1)}, \ldots, r_m^{(s)}, t^{[o]}) \quad \forall i \in [m] \setminus \{q\}$$

$$r_q^{(s+1)} = -\sum_{i \neq q} (r_i^{(s+1)})$$

Given the zero-mean constraint we imposed, it is essential to make sure that when updating the spatial frailty, the value of $r_q$ is adjusted accordingly to maintain the zero-mean constraint.

The full conditional distribution for each variable can be determined as follows:

1. $P(\beta^{(s+1)}|\tau^{2(s)}, \gamma^{(s)}, r^{(s)}, t^{[o]})$

$$P(\beta|\tau, \gamma, r, t^{[o]}) \propto P(\beta, \gamma, \tau^2, r, t^{[o]})$$

$$\propto \prod_{i=1}^{m} P(t_i^{[o]}|\beta, \gamma, r_i)P(\beta)$$

In the absence of an explicit form for the full conditional distribution, we employ rejection
sampling with the proposal density set to the probability density function (PDF) of the prior distribution for $\beta$ (i.e., $N(\beta_0, B_0)$). We explore various values of $\beta$ to gain an understanding of the upper bound, denoted as $M$, for the ratio between the target density and the proposal density (i.e., $\prod_{i=1}^m P(t_i^{[o]} | \beta, \gamma, r_i)$). It’s important to note that the likelihood $P(t_i^{[o]} | \beta, \gamma, r_i)$ is defined in Equations 3.2.1.

2. $P(\gamma_1^{(s+1)} | \beta^{(s+1)}, \gamma_2^{(s)}, \tau^2, r, t^{[o]}) \propto P(\beta, \gamma, \tau^2, r, t^{[o]})$

$$P(\gamma_1 | \beta, \gamma_2, \tau^2, r, t^{[o]}) \propto \prod_{i=1}^m P(t_i^{[o]} | \beta, \gamma, r_i) P(\gamma_1)$$

In cases where no closed-form solution can be derived for the full conditional distribution, we resort to rejection sampling once again, this time employing the PDF of the prior distribution for $\gamma_1$ (i.e., $N(\gamma_0, V_0)$) as the proposal density.

3. $P(\gamma_2^{(s+1)} | \beta^{(s+1)}, \gamma_1^{(s+1)}, \tau^2, r, t^{[o]}) \propto \prod_{i=1}^m P(t_i^{[o]} | \beta, \gamma, r_i) P(\gamma_2)$

In a similar fashion, we employ rejection sampling with the proposal density $P(\gamma_2)$ to address the absence of an analytical solution for the full conditional distribution of $\gamma_2$.

4. $P(\tau^{2(s+1)} | \beta^{(s+1)}, \gamma^{(s+1)}, r^{(s)}, t^{[o]})$

$$P(\tau^2 | \beta, \gamma, r, t^{[o]}) \propto P(\beta, \gamma, \tau^2, r, t^{[o]})$$

$$\propto P(r_{-q} | \tau^2) P(\tau^2)$$

$$\propto e^{-\frac{1}{2\tau^2} (r_{-q}^T (D_{-q} - \alpha r E_{-q}) r_{-q})} \tau^{2(m-1)} e^{-2(a + 1) \tau^{-2}} e^{-b \tau^{-2}}$$

$$\propto \tau^{-2(a + \frac{m-1}{2})} e^{-\frac{1}{2} (r_{-q}^T (D_{-q} - \alpha r E_{-q}) r_{-q}) \tau^{-2}}$$

$$\Rightarrow \tau^{-2(s+1)} \sim \Gamma(a + \frac{m - 1}{2}, b + \frac{1}{2} (r_{-q}^T (D_{-q} - \alpha r E_{-q}) r_{-q}))$$

Regarding the spatial precision parameter, $\tau^{-2}$, we benefit from a closed-form solution because
of the choice of a conjugate prior, allowing us to draw $\tau^{-2}$ from a gamma distribution.

5. $P(r_i^{(s+1)} | \beta^{(s+1)}, \gamma^{(s+1)}, \tau^2, r_1^{(s+1)}, \ldots, r_{i-1}^{(s+1)}, r_{i+1}^{(s)} \ldots, r_m^{(s)}, r_i^{[o]})$

We could perform updates on the spatial frailty for each subject individually. However, given the strong autocorrelation between samples due to spatial smoothness, we opted for blocked Gibbs sampling [28, 29] to reduce autocorrelation and expedite convergence. An intuitive way to group spatial frailties is based on their first-order neighbors, as the conditional autoregressive (CAR) prior relies solely on the spatial frailties of first-order neighbors.

However, an issue arises when leveraging blocked Gibbs sampling. As the full conditional distribution lacks an explicit form, we resort once again to rejection sampling. Nonetheless, when updating the spatial frailties of the central subject alongside its first-order neighbors, we encounter the "inefficiency of rejection sampling in high dimension" problem due to the curse of dimensionality, especially when the central subject has numerous neighbors. Our approach is to limit the number of first-order neighbors updated simultaneously. For instance, we limit to update spatial frailties for at most three subjects at a time (one central subject and two of its first-order neighbors). If the central subject has more than two first-order neighbors, our algorithm randomly selects two of them for updating.

Suppose our central subject is the $i^{th}$ subject, and we’re interested in a subset of its first-order neighbors ($N_s(i)$). Denote the corresponding selected second-order neighbors as $N_{s2}(i) := \{j : e_{j,k} = 1, \forall k \in N_s(i) \text{ and } j \neq i\}$. Let’s denote the union of them as $\tilde{N}_s(i) := \{N_s(i) \cup i\}$.

$$P(r_{\tilde{N}_s(i)} | \beta, \gamma, \tau^2, r_{N_s(i)} \setminus r_{\tilde{N}_s(i)}, r_i^{[o]}) \propto \prod_{j \in \tilde{N}_s(i)} P(r_j^{[o]} | \beta, \gamma, r_j) P(r_q^{[o]} | \beta, \gamma, r_q) P(r_{\tilde{N}_s(i)} \setminus \tilde{N}_s(i) \cup q, \tau^2)$$

$$\propto \prod_{j \in \tilde{N}_s(i)} P(r_j^{[o]} | \beta, \gamma, r_j) P(r_q^{[o]} | \beta, \gamma, r_q) P(r_{\tilde{N}_s(i)} \setminus \tilde{N}_s(i) \cup q, \tau^2)$$

In the second line, we apply Bayes’ theorem, and $N_{s3}(i) := \{N_{s2}(i) \cup \{N(i) \setminus N_s(i)\}\}$ represents

16
Figure 3.3: Subsets of neighbors introduced for updating the spatial frailties. Solid subjects’ spatial frailties are being updated given the spatial frailties of those marked by "X".

the union of the selected second-order neighbors and the unselected first-order neighbors. \( \tilde{\mathcal{N}}_{s_3}(i) := \{\tilde{\mathcal{N}}_s(i) \cup N_{s_3}(i)\} \) signifies the union of the central node, all of its first-order neighbors, and the selected second-order neighbors. The different subsets of neighbors introduced are illustrated in Figure 3.3.

Once again, the full conditional distribution lacks an explicit form. To address this, we utilize rejection sampling with the proposal density being \( P(r_{\tilde{\mathcal{N}}_s(i)}| r_{N_{s_3}(i)}, \tau^2) \). As an initial step in formulating the proposal density, it’s essential to recognize that \( P(r_{\tilde{\mathcal{N}}_s(i)}| r_{N_{s_3}(i)}, \tau^2) \propto P(r_{\tilde{\mathcal{N}}_{s_3}(i)}| \tau^2) \).

Based on \( r_{-q}| \tau^2 \sim N_{m-1}(0, \tau^2(D_{-q} - \alpha_r E_{-q})^{-1}) \) and the marginal of multivariate normal distribution, we derive the expression:

\[
 r_{\tilde{\mathcal{N}}_{s_3}(i)}| \tau^2 \sim N_{|\tilde{\mathcal{N}}_{s_3}(i)|}(0, \tau^2(D_{\tilde{\mathcal{N}}_{s_3}(i)} - \alpha_r E_{\tilde{\mathcal{N}}_{s_3}(i)})^{-1})
\]

Let \( \Sigma_s := \tau^2(D_{\tilde{\mathcal{N}}_{s_3}(i)} - \alpha_r E_{\tilde{\mathcal{N}}_{s_3}(i)})^{-1} \) denote the covariance matrix. Recall that \( \tilde{\mathcal{N}}_{s_3}(i) = \{\tilde{\mathcal{N}}_s(i) \cup \)
\(N_{s3}(i)\), we can arrange the spatial frailty and the covariance matrix as follows:

\[
\begin{bmatrix}
\tilde{r}_{N_{s3}(i)} \\
\tilde{r}_{N_{s3}(i)}
\end{bmatrix} =
\begin{bmatrix}
\Sigma_{s,11} & \Sigma_{s,13} \\
\Sigma_{s,31} & \Sigma_{s,33}
\end{bmatrix} = \Sigma_s
\]

Thus, the proposal density is,

\[
\tilde{r}_{N_{s3}(i)} | r_{N_{s3}(i)}, \tau^2 \sim N(\tilde{r}_{N_{s3}(i)})(\Sigma_{s,13} \Sigma_{s,33}^{-1} r_{N_{s3}(i)}, \Sigma_{s,11} - \Sigma_{s,13} \Sigma_{s,33}^{-1} \Sigma_{s,31})
\]

The steps of algorithm is summarized in Algorithm 1.

### 3.2.3 Posterior Inference

To deduce the censoring type indicator for censored subjects, the focus is on approximating the probabilities associated with left-censoring and right-censoring given that the subject is censored. These conditional probabilities are expressed using the survival function. For left-censoring, the probability is calculated as:

\[
P(t_i \leq T_1 | t_i^{[o]} = \text{NA}) = P(t_i < T_1 | t_i \notin (T_1, T_2)) = \frac{P(t_i < T_1)}{P(t_i \notin (T_1, T_2))} = \frac{1 - S_{\beta,\gamma,r_i}(T_1)}{1 - S_{\beta,\gamma,r_i}(T_1) + S_{\beta,\gamma,r_i}(T_2)}
\]

Similarly, for right-censoring:

\[
P(t_i \geq T_1 | t_i^{[o]} = \text{NA}) = \frac{1 - S_{\beta,\gamma,r_i}(T_2)}{1 - S_{\beta,\gamma,r_i}(T_1) + S_{\beta,\gamma,r_i}(T_2)}
\]

Once samples are obtained from the Gibbs sampler (Algorithm 1), the inference of the censoring type indicator is achieved through Monte Carlo approximation. The parameters and the
Algorithm 1: Gibbs Sampler for the Missing Censoring Type Indicator Problem

**Data:** \( X = \{ (v_i, t_i^{[o]}) \}_{i=1}^m \cup E \)

**Input:** \( P(t_i^{[o]} | \beta, \gamma, r_i) \) is defined in Equations 3.2.1
- \( P(\beta) \) is the PDF of \( N(\beta_0, B_0) \)
- \( P(\gamma) \) is the PDF of \( N(\gamma_0, V_0) \)
- \( P(r_i) \) is the PDF of \( N(\gamma_0, V_0) \)
- \( P(r_{N_{s3}(i)} | \Sigma_{s,13}^{-1} \Sigma_{s,33}, \Sigma_{s,11} - \Sigma_{s,13}^{-1} \Sigma_{s,31}) \)

\( r_i \leftarrow \begin{cases} i & \text{if } r_i \text{ has been updated at the } s^{th} \text{ iteration} \\ r_i^{(s-1)} & \text{else} \end{cases} \)

Randomly initialize \( \beta^{(0)}, \gamma^{(0)} \), \( \tau^{(0)}, r_1^{(0)}, \ldots, r_m^{(0)} \),

\( r_q^{(0)} \leftarrow - \sum_{i \neq q} (r_i^{(0)}) \); \( */ * \) for zero-mean constraint */

for iteration \( s = 1, \ldots, S \) do

1. Use rejection sampling to draw \( \beta^{(s)} \), with target density
   \( \propto \prod_{i=1}^m P(t_i^{[o]} | \beta^{(s)}, \gamma^{(s-1)}, r_i^{(s-1)}) P(\beta^{(s)}) \) and proposal density being \( P(\beta^{(s)}) \).

2. Use rejection sampling to draw \( \gamma_1^{(s)} \), with target density
   \( \propto \prod_{i=1}^m P(t_i^{[o]} | \beta^{(s)}, \gamma_1^{(s)}, \gamma_2^{(s-1)}, r_i^{(s-1)}) P(\gamma_1^{(s)}) \) and proposal density being \( P(\gamma_1^{(s)}) \).

3. Use rejection sampling to draw \( \gamma_2^{(s)} \), with target density
   \( \propto \prod_{i=1}^m P(t_i^{[o]} | \beta^{(s)}, \gamma^{(s)}, r_i^{(s-1)}) P(\gamma_2^{(s)}) \) and proposal density being \( P(\gamma_2^{(s)}) \).

4. Draw \( \tau^{-2(s)} \sim \Gamma(a_\tau + \frac{m-1}{2}, b_\tau + \frac{1}{2} (r_{(s-1)} - D_{(s-1)}) \).

5. Draw spatial frailties:
   (a) Randomly select a subset of central subjects \( C \) from \( \{1, \ldots, q - 1, q + 1, \ldots, m\} \);
   (b) for \( i \in C \) do

   Use rejection sampling to draw \( r_{N_{s3}(i)}^{(s)} \), with target density
   \( \propto \prod_{j \in N_{s}(i)} P(t_j^{[o]} | \beta^{(s)}, \gamma^{(s)}, r_{N_{s3}(i)}^{(s)}) P(t_j^{[o]} | \beta^{(s)}, \gamma^{(s)}, r_{N_{s3}(i)}^{(s)}) P(r_{N_{s3}(i)}^{(s)}) \) \( r_{N_{s3}(i)}^{(s)}, \tau^{2(s)} \) and proposal density being \( P(r_{N_{s3}(i)}^{(s)} | r_{N_{s3}(i)}^{(s)}, \tau^{2(s)}) \);

end

missing variable \( r \) in our model are collectively represented as \( \theta := (\beta, \gamma, \tau^2, r) \) If the Gibbs sampler samples, after discarding the burn-in period and thinning, are denoted as \( \{\theta^{(1)}, \ldots, \theta^{(T_s)}\} \), the Monte Carlo approximation for the conditional left- and right-censoring probabilities are computed
as follows:

\[
P(t_i \leq T_1 | t_i^{[o]} = NA) \approx \frac{1}{T_S} \sum_{s=1}^{T_S} \frac{1 - S_{\hat{g}(s), \gamma(s), \hat{r}(s)}(T_1)}{1 - S_{\hat{g}(s), \gamma(s), \hat{r}(s)}(T_1) + S_{\hat{g}(s), \gamma(s), \hat{r}(s)}(T_2)}
\]

\[
P(t_i \geq T_2 | t_i^{[o]} = NA) \approx \frac{1}{T_S} \sum_{s=1}^{T_S} \frac{S_{\hat{g}(s), \gamma(s), \hat{r}(s)}(T_2)}{1 - S_{\hat{g}(s), \gamma(s), \hat{r}(s)}(T_1) + S_{\hat{g}(s), \gamma(s), \hat{r}(s)}(T_2)}
\]

3.3 Hierarchical Model

3.3.1 Model Set-Up

We can extend our spatial survival model to a hierarchical model. Suppose there are a total of K groups. For the \(k^{th}\) group, there are \(J_k\) samples, and the \(j^{th}\) sample of the \(k^{th}\) group has \(m_{jk}\) subjects. On top of the sample-level parameters and subject-level spatial frailties introduced in
Section 3.2.1, the hierarchical model introduces group-level parameters as below,

**group-level** \((k = 1, \cdots K)\):

\[
\tilde{\beta}_k \sim N(\tilde{\beta}_0, B_{00}) \\
\tilde{\gamma}_{1,k} \sim N(\tilde{\gamma}_{1,0}, V_{00}) \\
\tilde{\gamma}_{2,k} \sim N(\tilde{\gamma}_{2,0}, V_{00}) \\
a_{\tau,k} \sim \Gamma(\tilde{a}_{\tau}, \tilde{b}_{\tau}) \\
b_{\tau,k} \sim \Gamma(\tilde{a}_{b}, \tilde{b}_b)
\]

**sample-level** \((j = 1, \cdots J_k)\):

\[
\beta_{jk} \sim N(\tilde{\beta}_k, B_0) \\
\gamma_{1,jk} \sim N(\tilde{\gamma}_{1,k}, V_{01}) \\
\gamma_{2,jk} \sim N(\tilde{\gamma}_{2,k}, V_{02}) \\
\tau_{jk}^{-2} \sim \Gamma(a_{\tau,k}, b_{\tau,k})
\]

**subject-level** \((i = 1, \cdots m_{jk})\):

\[
r_{m_{jk} \setminus jk} \sim N_{m_{jk} - 1}(0, \tau_{jk}^{2}(D_{(-j-k)} - \alpha_r E_{(-j-k)})^{-1}) \\
r_{q_{jk}} = - \sum_{i \neq q_{jk}} (r_{ijk})
\]

\[
P(t_{ijk}^{[o]} | \beta_{jk}, \gamma_{jk}, r_{ijk}) = \begin{cases} 
  f_{\beta_{jk}, \gamma_{jk}, r_{ijk}} (t_{ijk}^{[o]}) & \text{if } t_{ijk}^{[o]} \neq NA \\
  1 - S_{\beta_{jk}, \gamma_{jk}, r_{ijk}} (T_{1,jk}) + S_{i, \beta_{jk}, \gamma_{jk}} (T_{2,jk}) & \text{else}
\end{cases}
\]

\[
S_{\beta_{jk}, \gamma_{jk}, r_{ijk}} (t) = \frac{e^{-(r_{ijk} + \beta_{jk} v_{ij})} S_{\gamma_{jk}} (t)}{1 + (e^{-(r_{ijk} + \beta_{jk} v_{ij})} - 1) S_{\gamma_{jk}} (t)}
\]

\[
f_{\beta_{jk}, \gamma_{jk}, r_{ijk}} (t) = \frac{e^{-(r_{ijk} + \beta_{jk} v_{ij})} f_{\gamma_{jk}} (t)}{[1 + (e^{-(r_{ijk} + \beta_{jk} v_{ij})} - 1) S_{\gamma_{jk}} (t)]^2}
\]

The graphical model representation is illustrated in Figure 3.4. In comparison to Figure 3.2, it becomes evident how we extend our original model to the hierarchical model through the incorporation of an additional layer of group-level parameters.
3.3.2 Gibbs Sampler

The joint distribution of the latent and observed variables is:

\[
P(\tilde{\beta}, \tilde{\gamma}_1, \tilde{\gamma}_2, a_{\tau, k}, b_{\tau, k}, \beta, \gamma_1, \gamma_2, \tau^2, r, t) = \prod_{k=1}^{K} \prod_{j=1}^{J_k} \prod_{i=1}^{m_{jk}} P(t_{ijk} | \beta_{jk}, \gamma_{jk}, r_{ijk})
\]

\[
P(r_{m_{jk}} | q_{jk}) P(\tau^2_{jk} | a_{\tau, k}, b_{\tau, k}) P(a_{\tau, k}) P(b_{\tau, k})
\]

\[
P(\gamma_1_{jk} | \tilde{\gamma}_{1, k}) P(\gamma_2_{jk} | \tilde{\gamma}_{2, k}) P(\tilde{\gamma}_{2, k})
\]

\[
P(\beta_{jk} | \tilde{\beta}_k) P(\tilde{\beta}_k)
\]

The full conditional distribution for each variable can be determined in a similar manner as before:
**Group-level parameters**

1. \(P(\tilde{\beta}_k|\tilde{\gamma}_1,\tilde{\gamma}_2, a, b, \beta, \gamma_1, \gamma_2, \tau^2, r, t^{[o]}):\)

\[
P(\tilde{\beta}_k|\tilde{\gamma}_1,\tilde{\gamma}_2, a, b, \beta, \gamma_1, \gamma_2, \tau^2, r, t^{[o]}) \propto \prod_{j=1}^{J_k} P(\beta_{jk}|\tilde{\beta}_k) P(\tilde{\beta}_k)
\]

Since the conjugate prior of normal likelihood is normal, we can sample \(\tilde{\beta}_k\) from normal distribution as the following,

\[
\tilde{\beta}_k|\beta \sim N \left( \frac{1}{B_{00} + \frac{J_k}{B_0}} \left( \tilde{\beta}_{0} + \frac{\sum_{j=1}^{J_k} \beta_{jk}}{B_0} \right), \frac{1}{B_{00} + \frac{J_k}{B_0}} \right)
\]

2. \(P(\tilde{\gamma}_{1,k}|\tilde{\beta}, \tilde{\gamma}_2, a, b, \beta, \gamma_1, \gamma_2, \tau^2, r, t^{[o]}):\)

\[
P(\tilde{\gamma}_{1,k}|\tilde{\beta}, \tilde{\gamma}_2, a, b, \beta, \gamma_1, \gamma_2, \tau^2, r, t^{[o]}) \propto \prod_{j=1}^{J_k} P(\gamma_{1,jk}|\tilde{\gamma}_{1,k}) P(\tilde{\gamma}_{1,k})
\]

Again, leveraging the benefit of a conjugate prior, we can sample \(\tilde{\gamma}_{1,k}\) from normal distribution as the following,

\[
\tilde{\gamma}_{1,k}|\gamma_1 \sim N \left( \frac{1}{V_{00} + \frac{J_k}{V_0}} \left( \tilde{\gamma}_{1,0} + \frac{\sum_{j=1}^{J_k} \gamma_{1,jk}}{V_0} \right), \frac{1}{V_{00} + \frac{J_k}{V_0}} \right)
\]

3. Similar result for the full conditional distribution of \(\tilde{\gamma}_{2,k}\),

\[
\tilde{\gamma}_{2,k}|\gamma_2 \sim N \left( \frac{1}{V_{00} + \frac{J_k}{V_0}} \left( \tilde{\gamma}_{2,0} + \frac{\sum_{j=1}^{J_k} \gamma_{2,jk}}{V_0} \right), \frac{1}{V_{00} + \frac{J_k}{V_0}} \right)
\]
4. \( P(a_{\tau,k}|\bar{\beta}, \bar{\gamma}_1, \bar{\gamma}_2, b_\tau, \beta, \gamma_1, \gamma_2, \tau^2, r, t^{[o]}) \):

\[
P(a_{\tau,k}|\bar{\beta}, \bar{\gamma}_1, \bar{\gamma}_2, b_\tau, \beta, \gamma_1, \gamma_2, \tau^2, r, t^{[o]}) = P(a_{\tau,k}|b_\tau, \tau^2)
\]

\[
\propto \prod_{j=1}^{J_k} \frac{P(\tau_{jk}^2|a_{\tau,k}, b_{\tau,k}) P(a_{\tau,k})}{\tau_{jk}^{-2} \sim \Gamma(a_{\tau,k}, b_{\tau,k})} \Gamma(\tilde{a}_a, \tilde{b}_a)
\]

As the conditional lacks a closed-form solution, we again resort to rejection sampling with proposal density being \( P(a_{\tau,k}) \) (i.e. the PDF of \( \Gamma(\tilde{a}_a, \tilde{b}_a) \)).

5. \( P(b_{\tau,k}|\bar{\beta}, \bar{\gamma}_1, \bar{\gamma}_2, a_\tau, \beta, \gamma_1, \gamma_2, \tau^2, r, t^{[o]}) \):

\[
P(b_{\tau,k}|\bar{\beta}, \bar{\gamma}_1, \bar{\gamma}_2, a_\tau, \beta, \gamma_1, \gamma_2, \tau^2, r, t^{[o]}) \propto \prod_{j=1}^{J_k} \frac{P(\tau_{jk}^2|a_{\tau,k}, b_{\tau,k}) P(b_{\tau,k})}{\tau_{jk}^{-2} \sim \Gamma(a_{\tau,k}, b_{\tau,k})} \Gamma(\tilde{a}_b, \tilde{b}_b)
\]

Since the conjugate prior of Gamma likelihood with known shape is Gamma distribution, we can sample \( b_{\tau,k} \) from Gamma distribution as the following,

\[
b_{\tau,k}|a_\tau, \tau^2 \sim \Gamma(\tilde{a}_b + J_k \cdot a_{\tau,k}, \tilde{b}_b + \sum_{j=1}^{J_k} \tau_{jk}^{-2})
\]

**Sample-level parameters and subject-level spatial frailties**

The full conditional distributions for the sample-level parameters and the subject-level spatial frailties are essentially the same to those derived in Section 3.2.2.
Algorithm 2: Gibbs Sampler for the Missing Censoring Type Indicator Problem (Hierarchical Model)

Data: \( X = \{(v_{ijk}, t_{ijk}^0) : \forall i = 1, \cdots m_jk, j = 1, \cdots J_k, k = 1, \cdots K \} \cup E = \{E_{jk} : \forall j = 1, \cdots J_k, k = 1, \cdots K \} \)

/* Initialize randomly */

for group \( k = 1, \cdots K \) do

Randomly initialize \( \tilde{\beta}_k^{(0)}, \tilde{\gamma}_{1,k}^{(0)}, \tilde{\gamma}_{2,k}^{(0)}, a_{\tau,k}^{(0)}, b_{\tau,k}^{(0)} \);

for sample \( j = 1, \cdots J_k \) do

Randomly initialize \( \beta_{jk}^{(0)}, \gamma_{1,jk}^{(0)}, \gamma_{2,jk}^{(0)}, \tau_{jk}^{2(0)} \);

for subject \( i = 1, \cdots m_jk \) do

if \( i \neq q_{jk} \) then

Randomly initialize \( r_{ijk}^{(0)} \);

end

end

end

/* Update variables in each iteration */

for iteration \( s = 1, \cdots, S \) do

for group \( k = 1, \cdots K \) do

1. Draw \( \tilde{\beta}_k^{(s)} | \beta^{(s-1)} \sim N\left( \frac{1}{\nu_{00} + \nu_{01}^{(s-1)}} \left( \frac{\tilde{\beta}_0}{\nu_{00}} + \frac{\sum_{j=1}^{J_k} \beta_{jk}^{(s-1)}}{\nu_{01}} \right), \frac{1}{\nu_{00} + \nu_{01}^{(s-1)}} \right) \);

2. Draw \( \tilde{\gamma}_{1,k}^{(s)} | \gamma_{1}^{(s-1)} \sim N\left( \frac{1}{\nu_{00} + \nu_{01}^{(s-1)}} \left( \frac{\tilde{\gamma}_{1,0}}{\nu_{00}} + \frac{\sum_{j=1}^{J_k} \gamma_{1,jk}^{(s-1)}}{\nu_{01}} \right), \frac{1}{\nu_{00} + \nu_{01}^{(s-1)}} \right) \);

3. Draw \( \tilde{\gamma}_{2,k}^{(s)} | \gamma_{2}^{(s-1)} \sim N\left( \frac{1}{\nu_{00} + \nu_{01}^{(s-1)}} \left( \frac{\tilde{\gamma}_{2,0}}{\nu_{00}} + \frac{\sum_{j=1}^{J_k} \gamma_{2,jk}^{(s-1)}}{\nu_{01}} \right), \frac{1}{\nu_{00} + \nu_{01}^{(s-1)}} \right) \);

4. Use rejection sampling to draw \( a_{\tau,k}^{(s)} \) with target density \( \prod_{j=1}^{J_k} P(\tau_{jk}^{2(s-1)} | a_{\tau,k}^{(s)}, b_{\tau,k}^{(s-1)}) P(a_{\tau,k}^{(s)} | b_{\tau,k}^{(s-1)}) \) and proposal density being \( P(a_{\tau,k}^{(s)}) \);

5. Draw \( b_{\tau,k}^{(s)} | a_{\tau,k}^{(s)}, \tau_{jk}^{2(s-1)} \sim \Gamma(\tilde{a}_b + J_k \cdot a_{\tau,k}^{(s)}, \tilde{b}_b + \sum_{j=1}^{J_k} \tau_{jk}^{2(s-1)} \) ;

6. Update sample-level parameters and subject-level spatial frailties:

for sample \( j = 1, \cdots J_k \) do

Get updates on \( \beta_{jk}^{(s)}, \gamma_{1,jk}^{(s)}, \gamma_{2,jk}^{(s)}, \tau_{jk}^{2(s)}, r_{ijk}^{(0)} \) using Algorithm 1.

end

end

end
Embolisms, characterized by the blockage of xylem vessels with air, pose a critical challenge to the water transport system in plants, particularly angiosperms. In arid environments with limited soil water, heightened tension on the water column within xylem vessels can lead to the infiltration of air bubbles into these conduits. Subsequently, these air bubbles expand and disseminate throughout the xylem network, giving rise to air "embolisms." These embolisms impede the flow of water, potentially resulting in catastrophic hydraulic failure. Consequently, gaining insights into the spatiotemporal dynamics of embolism formation is indispensable for understanding how plants adapt to and manage extreme climatic conditions. Within this chapter, we elucidate the utility of our Gibbs Sampler in the context of embolism data.

4.1 Machine Learning Workflow for Embolism Event Detection

The relatively recent optical vulnerability technique (OV) [20, 21] offers a cost-effective approach to image embolism formation in plant leaves and stems. This method boasts a high temporal resolution, capturing images every 2 minutes. Moreover, research has demonstrated that the OV method yields estimates consistent with those obtained using standard hydraulic techniques [20, 21, 30].

Detection of an embolism event is achieved through an analysis of changes in reflectance between consecutive images in OV technique. As xylem vessels transition from water-filled to air-filled states, they shift from a translucent (typically dark) appearance to a reflective (white) one. By computing the differences between two sequential images, embolism events can be extracted.

The current method of extracting embolism events from images demands significant time and labor for large-scale optical vulnerability image datasets. While open-source image processing
programs such as ImageJ [31] have expedited some aspects of the process, the extraction procedure still relies on human involvement. Annotators must manually review each differentiated image, differentiate true embolism events from spurious artifacts, and pinpoint areas in an image where embolism occurs. As human expertise is a pivotal component of this process, the quality of embolism extraction may vary substantially from one annotator to another.

With the goal of developing an AI-assisted embolism event detection workflow, our primary metric of interest is the reduction of false negatives. We are less concerned about the relatively high occurrence of false positives, as long as we can retain potential embolism candidates. These results would remain valuable for mitigating the human detection workload. False positives can be subsequently minimized during the final phase of human filtering.

In this section, we introduce a machine learning (ML) workflow that enhances the detection of embolism events within high-resolution images, with a particular focus on the task of distinguishing embolism events from background noise. This AI-assisted event detection workflow not only accelerates the detection process but also ensures the consistency and quality of labeled data. Utilizing manually labeled stem images, we train an ML workflow for embolism detection, generating pixel-wise embolism predictions for three evergreen wet tropical forest angiosperm tree species: *Alchornea latifolia*, *Casearia arborea*, and *Inga laurina*. Our data originates from El Yunque National Forest in Puerto Rico, where we conducted extensive imaging using the optical vulnerability technique. The principal objective is to detect embolism events within these fine-resolution stem images, and we subsequently evaluate our results through image-level accuracy and sensitivity.

4.1.1 Noises

The process of embolism extraction needs to address various sources of noise, which include:

1. **Shifting:** The stem exhibits consistent and sudden shifting, primarily attributable to gel movement. The magnitude of this shift varies, ranging from minor, inconsequential movements to more substantial shifts that can mislead observers into identifying numerous embolism events when examining the binarized difference image, as a consequence of the image
2. Bark: In some instances, the stem’s bark is not entirely removed before image acquisition. Along with the shifting effect, one might see dark lines in the difference image near the boundary of stems (Figure 4.1a).

3. Shrinking: Over time, the stem undergoes a gradual reduction in size due to dehydration, leading to complications in image analysis.

4. Browning: The stem’s color transforms as it dehydrates, transitioning from a green hue to various shades of brown, further complicating the assessment process.

5. Color Variability: The stem’s coloration in the images is not consistently "green stem against a red background," with instances where the stem appears light yellow against a dark brown background.
Figure 4.2: The machine learning workflow for detecting embolism events using optical vulnerability stem image data.

6. **Plastic Cover**: The curvature contour of the petri dish becomes visible due to shifting (Figure 4.1b), introducing an additional layer of complexity for embolism identification.

7. **Bubbles**: Gel bubbles exhibit mobility during the dehydration process, contributing to the presence of noise that requires careful handling during image analysis (Figure 4.1c).

4.1.2 Machine Learning Workflow

The proposed machine learning workflow comprises four stages, as illustrated in Figure 4.2.

1. **Differencing and Binarizing**

   The process involves transforming raw images into grayscale images and then computing the difference between consecutive images, serving as the initial stage for embolism event detection. Subsequently, this difference image is binarized by assigning a value of 1 to all positive pixel values, denoting regions of change between two consecutive grayscale images. A value of 0, on the other hand, signifies the absence of change at a particular pixel between the two images. It is noteworthy that these initial stages closely resemble the procedures employed when utilizing ImageJ. However, it is important to emphasize that the subsequent stages extend beyond the capabilities of ImageJ.
2. **Foreground Background Segmentation**

To combat the shifting noise, we would like to extract stem from the images. This can be done by shifting the user-provided stem mask using correlation between consecutive images, as correlation between consecutive images can be used to identify changes in position. If the user can’t provide a mask for stem, Otsu’s method [32] can be used on the grayscaled raw images to automatically extract areas of stem. The concept of Otsu’s method is to minimize the within-class variance. First, it computes the histogram of pixel intensity level for an input image. Then, it finds the optimal thresholding values that minimizes the within-class intensity variance by trying out many possible thresholding values. Otsu’s method provides an adaptive threshold for each image, meaning the threshold is not fixed over time. This a good consideration since stem becomes browner as time goes by because of dehydration. However, for later images, pixel level intensity is not enough for foreground background segmentation, as the stem color becomes so dark that it’s even hard to distinguish the foreground from background even through human eyes.

3. **Poor Quality Detection**

In cases where an image contains an excessive number of circular elements, which could potentially represent bubbles, embolism events, shifting artifacts, or the presence of a plastic cover, it is designated as "poor quality." The detection of circles in these images is accomplished using Hough transformations [33]. As a consequence, images categorized as "poor quality" are presently presumed to lack embolism events. Additionally, the algorithm generates a list of indices corresponding to the images flagged as poor quality.

4. **De-noising**

During the de-noising stage, a variety of distinct techniques are employed, forming a cascade of de-noising steps.

- **Remove Extreme Candidates by Basic Shape Analysis**

  Mitigating false positives resulting from shifting artifacts, we use morphological image
processing techniques to initially identify a set of potential embolism candidates. In the event that any connected component exhibits excessive width, substantial area, and high density, classifying the image as "poor quality."

- **Remove Reappearing Candidates**
  To mitigate the inclusion of false positives attributed to plastic covers or bubbles, a criterion is applied: recurrent appearances are not counted as embolism events. This criterion is operationalized by setting a threshold on the frequency of non-overlapping sections (utilizing a rolling window approach). Any instances surpassing this threshold are not considered as embolism events.

- **Separating Weak and Strong Candidates**
  Subsequent to the removal of recurrent candidates, our analysis proceeds with the elimination of outlier candidates based on shape analysis. Specifically, we apply two criteria for the identification and subsequent removal of connected components:
  Firstly, connected components characterized by low density and exhibiting either small size or limited height are considered for removal.
  Secondly, we identify connected components as candidates for removal if they demonstrate excessive size, width, or if their width-to-length ratio is disproportionately large.
  In the subsequent step, we categorize the candidates into two groups, distinguishing between weak and strong candidates. Weak candidates are defined as those images containing exclusively "small or short but high-density" events. Conversely, any images that do not meet this criterion are categorized as strong candidates.

- **(Optional) Dual-Input Convolutional Neural Network (CNN):**
  Typically, the number of strong candidates significantly outweighs that of weak candidates, with strong candidate images constituting approximately 70 percent of the remaining candidates. Given the relatively limited information contained in the weak candidates and their smaller proportion in comparison to the strong candidates, our focus in this phase is exclusively directed towards the analysis of strong candidates.
Table 4.1: Data Summary.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Alchornea latifolia</em></td>
<td>ALCLAT</td>
<td>1</td>
<td>1405</td>
<td>80</td>
</tr>
<tr>
<td><em>Casearia arborea</em></td>
<td>CASARB</td>
<td>2</td>
<td>1393</td>
<td>43</td>
</tr>
<tr>
<td><em>Inga laurina</em></td>
<td>INGLAU</td>
<td>3</td>
<td>1495.33</td>
<td>11.33</td>
</tr>
</tbody>
</table>

To further refine the identification of false positives among the strong candidates, we employ a dual-input Convolutional Neural Network (CNN). The term "dual" in this context signifies the utilization of both the strong candidate images and their corresponding binarized difference images obtained during the initial stage. While this approach shows promise in terms of reducing false positives, it is important to acknowledge that it may lead to some reduction in sensitivity (false negative).

- Remove Small Candidates

In our continued efforts to curtail false positives, we implement an additional strategy at the image level. Specifically, if the proportion of embolized pixels in an image is deemed excessively small, the entire image is treated as devoid of any embolism events. This approach stands in contrast to the earlier steps involving shape analysis, which primarily operated at the component level.

4.1.3 Stem Data

Table 4.1 provides a comprehensive listing of the species utilized in this study, including their corresponding species codes employed in other tables. It further specifies the number of leaves that were subject to measurement for each species, along with the average count of images captured per leaf. Additionally, the table furnishes the average count of images featuring embolism events per leaf.

Notably, it is imperative to observe that the proportion of images exhibiting embolism events is relatively low, signifying that the detection of such events is indeed a rare occurrence.
4.1.4 Implementation Details

In Step 3, known as the "Poor Quality Detection" phase, images are flagged as poor quality if the detected circles occupy more than 10% of the stem area.

Within the framework of Step 4, denoted as the "De-noising" phase, a set of parameter values has been determined through a process of trial and error.

In the "Remove Extreme Candidates by Basic Shape Analysis" segment, connected components are subject to elimination if their characteristics exceed certain thresholds. Specifically, components with a width greater than 0.85 of the stem width, an area exceeding 0.2 of the stem area, and a density surpassing 0.3 of the bounding box of the component are removed from consideration.

Moving on to the "Remove Reappearing Candidates" step, a rolling window encompassing 200 images is employed, with a predetermined threshold value set at 0.05. The criterion for flagging pixel locations as devoid of embolism is if a particular location exhibits a reappearing candidate in 10 images within a consecutive window of 200 images.

Lastly, the "Separating Weak and Strong Candidates" stage involves the retention of connected components meeting specific criteria. Such components are preserved if their height exceeds 49 pixels, their area falls within the range of 1000 to 75000 pixels, their width is in the range of 25 to 200 pixels, and their width is less than their height. Additionally, images containing components with heights between 40 and 49 pixels or areas between 700 and 1000 pixels are categorized as weak candidates.

The dual-input CNN first reshapes input images into size of 256 × 256. And it uses soft-max activation layer along with categorical crossentropy loss function. It’s trained with Adam optimizer [34], for 500 epochs, with batch size of 50. We used 80% of the data for training and 20% of the data for testing.
Table 4.2: Performance of the proposed machine learning workflow using different ways of foreground background segmentation.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ALCLAT</td>
<td>83.75%</td>
<td>88.32%</td>
<td>82.5%</td>
<td>90.21%</td>
</tr>
<tr>
<td>CASARB</td>
<td>88.13%</td>
<td>93.76%</td>
<td>88.13%</td>
<td>92.9%</td>
</tr>
<tr>
<td>INGLAU</td>
<td>87.53%</td>
<td>95.09%</td>
<td>87.53%</td>
<td>93.79%</td>
</tr>
<tr>
<td>Average</td>
<td>86.47%</td>
<td>92.39%</td>
<td>86.05%</td>
<td>92.3%</td>
</tr>
</tbody>
</table>

4.1.5 Performance Evaluation

Given our primary focus on reducing false negatives, our evaluation metrics extend beyond accuracy and include sensitivity, which is defined as the ratio of true positives to the sum of true positives and false negatives.

Upon reviewing the results presented in Table 4.2, it becomes evident that the overall performance remains consistent, regardless of whether a user-provided stem mask (default) or Otsu’s method is employed for foreground background segmentation. However, it is noteworthy that the performance of the former approach is significantly contingent on the quality of the user-provided stem mask. Consequently, we recommend the adoption of Otsu’s method for the automatic segmentation of the foreground stem from the background, as it offers a more robust and consistent solution.

We provide a distinct report on the performance of the dual-input CNN to facilitate users in making informed decisions. The dual-input CNN yields a testing sensitivity of 70.8% and an accuracy of 86.4%. This separation of results allows users to make choices tailored to their sensitivity to potential increases in false negatives, providing them with the flexibility to opt out of incorporating the dual-input CNN in their workflow if deemed necessary.

Given its promising performance metrics, with sensitivity approximately at 86% and accuracy around 92%, our proposed machine learning workflow emerges as a highly valuable tool for embolism event detection. This proficiency in data handling not only signifies the potential efficacy
of our workflow but also lays the foundation for the subsequent application of our Gibbs Sampler to embolism data. The amelioration of event detection burdens enables the seamless application of our Gibbs sampler to delve into the spatiotemporal dynamics governing embolism datasets, particularly under the influence of a missing censoring type indicator.

4.2 Exploratory Data Analysis

In this section, our focus is on elucidating the limitations stemming from the application of the spatial survival model to embolism data without addressing the issue of the missing censoring type indicator. Given that the predominant proportion of censored events within the embolism data aligns with the right-censoring category, the spatial survival model presented in this section makes the simplifying assumption of right-censoring.

4.2.1 Survival Model for Embolism Data: A Rationale

The appropriateness of applying a survival model to embolism data is justified by the intrinsic nature of the dataset. The primary focus in embolism data analysis is discerning the timing of embolism occurrences. Survival analysis, a well-established approach in medical literature [10, 12], proves pertinent for such investigations, particularly when analyzing time-to-event data, where the event of interest is the occurrence of embolism.

Moreover, the analogy that once a vessel conduit undergoes embolization, it remains in that state akin to the ‘death’ of a vessel conduit, aligns seamlessly with the principles of survival analysis commonly used in the medical literature. This alignment underscores the appropriateness of employing survival modeling techniques to effectively capture and interpret the temporal dynamics inherent in embolism datasets.

The prevalent method for summarizing xylem vulnerability data, acquired through standard hydraulic techniques, involves the utilization of "vulnerability curves." Notably, in instances where data is acquired through the optical vulnerability technique, studies have demonstrated that "optical vulnerability curves" offer a suitable approximation for these curves [20, 21, 30].
In reinforcing our approach, we establish a meaningful connection between survival functions and the commonly employed vulnerability curves. This connection serves to provide assurance that our proposed survival model is firmly grounded in the well-established framework of xylem analysis, further enhancing the applicability of our methodology.

The connection between the percentage loss of xylem conductance (PLC) and water potential, often referred to as pressure, is typically delineated through a vulnerability curve (VC). Over time, water potential tends to become more negative, with the reference point being that a fully hydrated branch registers a water potential of 0. Concurrently, PLC values progressively ascend from 0 to 1. To construct a vulnerability curve, it is necessary to acquire measurements of hydraulic conductivity \( K_i \) at various water potentials \( \Psi_i \), along with the determination of the maximum conductivity \( K_{max} \) for each branch. Assuming we have a total of \( n_{VC} \) pairs of measured hydraulic conductivity and water potentials \( \{K_i, \Psi_i\}_{i=1}^{n_{VC}} \), PLC values can be computed as follows:

\[
PLC(K_i, K_{max}) = 1 - \frac{K_i}{K_{max}} \times 100\% \tag{4.1}
\]

A common way of representing the relationship between conductance \( K_i \) and water potential \( \Psi_i \) is through Weibull formulation with scale \( \beta > 0 \) and shape \( \alpha > 0 \)[26, 27]:

\[
\frac{K_i}{K_{max}} = e^{-\left(\frac{\Psi_i}{\beta}\right)^\alpha} \tag{4.2}
\]

As for survival analysis, suppose we denote the density that survival time \( T \) follows as \( f(t) \), then the survival function, the probability an object of interest has survived till time \( t \), is defined as:

\[
S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du \tag{4.3}
\]

If we assume that survival time \( T \) follows a Weibull distribution also with scale \( \beta > 0 \) and
shape ($\alpha > 0$), we can write out the density function as,

$$f(t) = \frac{\alpha}{\beta} (\frac{t}{\beta})^{\alpha-1} e^{-\left(\frac{t}{\beta}\right)^{\alpha}}$$  \hspace{1cm} (4.4)

And using the definition from Equation (4.3), the parametric estimator of the survival function can be easily derived to be as,

$$S(t) = e^{-\left(\frac{t}{\beta}\right)^{\alpha}}$$  \hspace{1cm} (4.5)

Essentially, Equations (4.2) and (4.5) are the are identical when time in the survival function is replaced by the negative of water potential. This replacement is logical, as water potential becomes increasingly negative as time elapses. Consequently, in accordance with the definition of PLC, the vulnerability curve (VC) derived from the Weibull function would correspond to the complement of the survival function using the Weibull distribution ($PLC(\Psi_i) = 1 - S(\Psi_i)$). The equivalence provides assurance that the survival analysis is firmly grounded in the well-established principles of xylem analysis.

Furthermore, it is worth noting that the median survival time derived from a survival curve is determined as the smallest survival time at which the survivor function equals or falls below 0.5. Notably, this aligns with the concept of $P_{50}$ for a vulnerability curve, defined as the water potential at which PLC reaches 50%. The parallels in the frequently employed summary statistics for both vulnerability curves and survival curves further underscore the rationale for employing survival modeling in the study of embolism formation. Furthermore, adopting the perspective of survival analysis enables us to extend our modeling efforts from "temporal" information to encompass the broader "spatio-temporal" context.

4.2.2 Leaf Data

The embolism leaf data [35] for this and the following section was gathered from eight prevalent evergreen tree species within the Luquillo Experimental Forest (LEF) situated in the El Yunque National Forest (EYNF) in northeastern Puerto Rico. The data collection process involved the uti-
Table 4.3: List of species used in this study, code used in figures, their families, and number of leafs measured per species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Code</th>
<th>Family</th>
<th>Num. leafs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alchornea latifolia</td>
<td>ALCLAT</td>
<td>Euphorbiaceae</td>
<td>4</td>
</tr>
<tr>
<td>Casearia arborea</td>
<td>CASARB</td>
<td>Flacourtiaee</td>
<td>8</td>
</tr>
<tr>
<td>Cecropia schreberiana</td>
<td>CEC SCH</td>
<td>Moraceae</td>
<td>7</td>
</tr>
<tr>
<td>Drypetes glauca</td>
<td>DRYGLA</td>
<td>Euphorbiaceae</td>
<td>7</td>
</tr>
<tr>
<td>Inga laurina</td>
<td>INGLAU</td>
<td>Fabaceae</td>
<td>6</td>
</tr>
<tr>
<td>Ocotea leucoxylon</td>
<td>OCOLEU</td>
<td>Lauraceae</td>
<td>5</td>
</tr>
<tr>
<td>Sloanea berteroana</td>
<td>SLOBER</td>
<td>Elaeocarpaceae</td>
<td>8</td>
</tr>
<tr>
<td>Tabebuia heterophylla</td>
<td>TABHET</td>
<td>Bignonaceae</td>
<td>8</td>
</tr>
</tbody>
</table>

lization of the optical vulnerability technique, as described in [20]. For each species, data was collected from a varying number of leaves, each originating from a distinct tree. The number of leaves collected for each species ranged from 4 to 8, as specified in Table 4.3.

The dataset comprises raw leaf images, manually extracted embolism images, and pairs of measured water potential and time. For each leaf, approximately 1500 raw images were captured at 2-minute intervals. To prepare the data for our proposed spatial survival model, several crucial data processing steps are required. These steps involve the transformation of water potential, the extraction of vein segments as the unit of analysis, the computation of vein thickness for each vein segment, the construction of a venation network to capture spatial dependencies between vein segments, and the assessment of whether an embolism event occurred for each vein segment.

**Water Potential Transformation**

For the purposes of our spatial survival model and the interests of phytologists, we intend to use water potential as the temporal variable. It’s important to note that water potential measurements were taken at 10-minute intervals, which doesn’t align with the image capture frequency. To address this discrepancy, we applied a monotone polynomial fitting method based on [36] to the water potential measurements, resulting in what we refer to as "transformed water potential."
The application of this monotone polynomial fitting ensures that the transformed water potential remains non-increasing as time progresses, thereby mitigating the effects of oscillatory noise in the measured water potential. This noise is primarily attributed to measurement errors in the psychrometers. We selected a degree of 3 for the monotone polynomial to strike a balance between fitting accuracy and preventing overfitting. This is especially important in cases where there are stepwise decreases in water potential due to the oscillatory noise associated with the psychrometers. Figure 4.3 demonstrates how a higher degree of monotone polynomial can lead to overfitting, especially when attempting to capture oscillatory noise. The transformed water potential, obtained through this procedure, allows us to assign a corresponding water potential value to each image through interpolation. For the sake of simplicity, later in this chapter, we will refer to the "transformed water potential" as simply "water potential."

Vein Segment Processing

Given the reasoning outlined in Section 4.2.1, our approach involves treating each "xylem vessel" (or "xylem conduit") as a subject in our survival models, since it’s the phytological unit
associated with embolism. Unfortunately, due to the inherent limitations in the resolution of the available images, we can only approximate these vessels using "vein segments".

To identify the veins, we utilized phenoVein [37], a semi-automated leaf vein segmentation software. The output of this software marked the positions of branching and ending points for each vessel in the sample. Using these markers, we defined a vein segment as the section of a vein located between two branching points or between one branching point and one ending point (see Figure 4.4). It’s worth noting that a single xylem conduit could be separated into multiple vein segments, which represents the best approximation possible given the current resolution of available images.

While we have approximated xylem conduits using vein segments due to current image resolution constraints, future advancements in imaging technology may facilitate more precise modeling. It’s worth noting that one can easily adapt our spatial survival model for xylem conduits.

**Vein Thickness Measurements**

To assess the significance of vein thickness in embolism formation, we employed binarized images to estimate the thickness of each vein segment. An intuitive method for estimating the segment’s thickness involves first identifying the medial axis (or the skeleton) of the vein segment and then calculating the width of the vein segment, which should be perpendicular to the medial axis. The medial axis was determined using the distance transform [38]. The distance transform provides the distance from each vein pixel to the nearest background pixel for each vein pixel, and the vein pixels with higher distance transform values constitute the medial axis. Assuming the symmetrical geometry of veins, the vein thickness is equivalent to twice the maximum pixel-level distance transform value of the medial axis (see Figure 4.5a-c).

Upon acquiring the thickness information for each vein segment, we can integrate this data into our spatial survival model by including vein thickness as a covariate (e.g., $v_i$ for the $i^{th}$ vein segment). Consequently, the covariate coefficient $\beta$, interpreted as the vein thickness coefficient, will capture the influence of vein thickness on embolism formation.
Figure 4.4: An example for vein segment processing. (a) Unprocessed image of a *Ocotea leucoxy- lon* leaf. (b) Binarized image derived from the first image using phenoVein. (c) Locations of the branching points and ending points are highlighted by yellow and red markers respectively through phenoVein. (d) Visualization of embolism events during desiccation in the observation period (i.e., increases in water potential, wp).
Venation Network Construction

To account for the impact of spatial connectivity within the venation network on the spatiotemporal propagation of embolisms, we established a definition for neighbors. Specifically, neighbors were defined as any two vein segments that were connected in the image obtained from phenoVein (as illustrated in Figure 4.4c). Utilizing this neighbor definition, we were able to create the adjacency matrix denoted as $E$, which encodes the spatial dependence between vein segments.

Embolism Events

To map embolized pixels from manually extracted embolism images to embolized vein segments, we employ the following strategy:

1. Initially, we consider all vein segments that have at least one embolized pixel to be "candidates."

2. We define the "overlapping percentage" as the intersection between embolized pixels and the pixels of a vein segment candidate, divided by the number of embolized pixels. This
percentage represents the proportion of embolized pixels belonging to a vein segment candidate.

3. Our approach involves iteratively removing candidates with overlapping percentages below 5% in ascending order of their overlapping percentages. The constraint here is that the total overlapping percentage from the remaining candidates must add up to at least 90%.

4. The final candidates that remain after this process are treated as vein segments that have experienced embolism events.

Our primary focus is the timing and location of these embolism events. To address the "where" aspect of this inquiry, we harnessed the spatially neighboring information, as established during the construction of the vein network. Meanwhile, to tackle the "when" dimension of this question, we leveraged the water potential associated with each embolism event. The rationale behind employing water potential for the "when" dimension lies in its implicit ability to signify the duration required for an embolism event to occur, essentially representing "time to event". This inference arises because water potential decreases over time. Using the spatial neighboring information and water potential of embolism events, we turned to our spatial survival model to unravel the spatiotemporal dynamics of these spatially referenced time-to-event embolism data.

4.2.3 Limitation of Only Considering Right-censoring

Comparing the vein segments in the raw image and the observed embolism events in Figures 4.4, it is evident that numerous vein did not yield observations of embolism events during the designated observation period. Given the observation that thicker veins tend to experience embolism earlier[20, 21, 22, 23, 24], an inference can be drawn that there may be a limited occurrence of left-censoring concomitant with a substantial prevalence of right-censoring among the censored vein segments.

To elucidate this, Figure 4.6 serves as an illustrative example where left-censoring is suspected to have occurred at the thicker midvein. However, if we used a spatial survival model that only
Figure 4.6: Example of left-censoring in a *Casearia arborea* leaf. (a) Observed embolism events. Dashed circle highlights where left-censoring is suspected to occur. The color red indicates that the vein segment embolizes earlier, whereas purple signifies a later embolism occurrence. (b) Posterior mean of log odds ratio if we only assume right-censoring. Notice that the circled midvein is predicted to be right-censored.
assumed right-censoring, it would estimate the censored midvein to embolize at a later time. This outcome aligns with expectations, as the model did not account for the left-censoring possibility and in its model formulation.

4.3 Application of the Gibbs

To gauge the accuracy of our model, we conducted a thorough evaluation encompassing a series of rigorous analyses and comparisons. Our findings unequivocally demonstrate the model's reliability in handling challenging scenarios characterized by temporal uncertainties associated with missing data.

Our initial assessment involved evaluating the performance of our model through a meticulously designed simulated dataset, carefully crafted to replicate the intricate scenarios encountered in our real-world applications. This simulated dataset served as a valuable tool for quantifying the model’s ability to infer the missing censoring type indicator, particularly in the absence of ground-truth information for the real-world embolism data. Subsequently, we applied our model to real-world datasets from ecological contexts, specifically the embolism leaf data detailed in Section 4.2.2. While the direct evaluation of our model’s performance in inferring the missing censoring type indicator proved challenging due to the absence of a straightforward ground truth, we provided visualizations of our estimates and showcased the efficacy of our Gibbs sampler in addressing the issue presented in Section 4.2.3, yielding results in line with expectations. Additionally, we illustrated that the estimates generated by our model not only align with established phytological findings but also have the potential to uncover novel insights.

4.3.1 Simulation: Parameter and Censorship Type Indicator Recovery

Simulated Data Description

To assess the performance of our Gibbs sampler and verify its functionality under scenarios resembling the model assumptions, we generated a simulated dataset that aligns with the model specifications outlined in Section 3.2.1. We generated a simulated dataset by first sampling pa-
Table 4.4: Results for predicting the missing censoring type indicator. In the confusion matrix, left-censored, right-censored, and uncensored are abbreviated as "l", "r", and "u" respectively. The overall accuracy encompasses both censored and uncensored cases, whereas censored accuracy specifically considers only censored cases.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Truth</th>
<th>Prediction</th>
<th>Truth</th>
<th>Prediction</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>l r u</td>
<td>l r u</td>
<td>l r u</td>
<td>l r u</td>
</tr>
<tr>
<td>Confusion Matrix</td>
<td></td>
<td>4 0 0</td>
<td>l 3 0 0</td>
<td>l 0 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 72 0</td>
<td>r 3 72 0</td>
<td>r 6 72 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 0 224</td>
<td>u 0 0 224</td>
<td>u 0 0 224</td>
<td></td>
</tr>
<tr>
<td>Overall Acc.</td>
<td>0.9934</td>
<td>0.9901</td>
<td>0.9801</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cens. Acc.</td>
<td>0.9744</td>
<td>0.9615</td>
<td>0.9231</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

rameters along with the unobserved spatial frailties (i.e., \( \theta \)) from the model described in Section 3.2.1. Subsequently, we simulated the events using the \textit{simsurv} package [39]. To replicate the ve-
nation network found in real-world leaf embolism data, we utilized the adjacency matrix and vein thickness information from a leaf sample of the species \textit{Drypetes glauca} when simulating events. Furthermore, to emulate the imbalanced nature of different censoring type observed in real-world leaf embolism data, as discussed in Section 4.2, we tailored our observation period to include more instances of right-censoring and fewer instances of left-censoring. In the end, we obtained a sim-
ulated dataset that closely adheres to the assumed data-generating mechanism of our model. This simulated dataset incorporates a venation network structure akin to that observed in real-world datasets, and it captures the imbalanced distribution of different censoring types, mirroring the characteristics observed in actual real-world datasets.

In our simulated dataset, we considered a total of 302 vein segments (i.e., \( m = 302 \)). Analyz-
ing the confusion matrix in Table 4.4 reveals that the simulated data exhibits a higher prevalence of right-censored instances compared to left-censored instances (i.e., \( 72 > 6 \)). This intentional imbalance aligns with the observed distribution of different censoring types in the real-world data, further enhancing the simulation’s fidelity to the complexities inherent in practical scenarios.

The hyperparameters were set as specified in Table 4.5. The Gibbs sampler was executed 3
Table 4.5: The hyperparameter values for the simulated dataset.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>7.5</td>
</tr>
<tr>
<td>$B_0$</td>
<td>17.37</td>
</tr>
<tr>
<td>$\gamma_{0,1}$</td>
<td>-6</td>
</tr>
<tr>
<td>$V_{0,1}$</td>
<td>2.5</td>
</tr>
<tr>
<td>$V_{0,2}$</td>
<td>1.5</td>
</tr>
<tr>
<td>$a_\tau$</td>
<td>6.0</td>
</tr>
<tr>
<td>$b_\tau$</td>
<td>30.0</td>
</tr>
<tr>
<td>$a_r$</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Table 4.6: The simulated dataset’s true parameter values. The posterior mean and 90% interval obtained through our proposed Gibbs sampler and through the model only assumes right-censoring.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>Gibbs Mean (90% Interval)</th>
<th>Right Mean (90% Interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>3</td>
<td>2.96 (-0.55, 4.65)</td>
<td>2.97 (2.28, 3.78)</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-6.4</td>
<td>-6.42 (-7.08, -5.91)</td>
<td>-6.29 (-6.38, -6.23)</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>2</td>
<td>1.69 (0.81, 2)</td>
<td>1.8 (1.71, 1.91)</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>10</td>
<td>5.29 (3.42, 8.18)</td>
<td>3.38 (2.03, 5.25)</td>
</tr>
</tbody>
</table>

Performance Evaluation

Our initial evaluation of the Gibbs sampler’s performance primarily centered on parameter recovery. The results, presented in Table 4.6, demonstrate that the posterior means generated by our Gibbs sampler closely approximate the true values for most parameters, except for the spatial frailty variance ($\tau^2$). Recovering $\tau^2$ poses a greater challenge due to an additional layer of uncertainty introduced at the "vein segment" level. Unlike other parameters, $\tau^2$ contends with
uncertainty arising from the spatial frailty of each vein segment, adding complexity to its recovery. Further details on the Gibbs sampler’s improved recovery of $\tau^2$ in scenarios with more events per vein segment, will be discussed in the upcoming empirical bias study (Section 4.3.3).

Comparing our Gibbs sampler with the right-censoring model, it performs similarly for $\beta$, $\gamma_1$, and $\gamma_2$, but outperforms significantly in estimating the challenging spatial frailty variance parameter $\tau^2$. The $\tau^2$ estimation of the right-censoring model is notably underestimated. While the right-censoring model appears less uncertain, with a narrower 90% interval, it may fail to encompass the true parameter values, indicating a potential for higher bias.

We further assessed the performance of our Gibbs sampler by evaluating its ability to infer the missing censoring type indicator. The comparison between the Gibbs sampler and the oracle, where true parameter values are known, is presented in Table 4.4. Notably, even with knowledge of the true parameter values, full recovery of the missing censoring type indicator may be challenging (overall accuracy is less than 1). Nevertheless, our results closely approximate those of the oracle, demonstrating high overall accuracy and censored accuracy. Particularly noteworthy is the substantial improvement in censored accuracy when comparing our Gibbs sampler compared to the model that only assumes right-censoring. The improvement in the Gibbs sampler’s performance indicates its effectiveness in successfully inferring the missing censoring type indicator.

4.3.2 Simulation: Robustness Analysis

In this section, we investigate the performance of our proposed method when the assumptions of the model deviate from the actual data generating mechanism. Previously, we demonstrated superior performance of our method compared to the right-censoring model on a dataset containing both left- and right-censoring. Now, we explore the reverse scenario where the data does not exhibit any left-censoring but exclusively presents right-censoring
**Simulated Data Description**

To address this inquiry, we simulated a dataset without left-censoring and evaluated through parameter recovery and censoring type indicator recovery. As in Section 4.3.1, we considered a total of 302 vein segments (i.e., \( m = 302 \)). The confusion matrix in Table 4.7 reveals that the simulated data contains only 64 right-censored instances and no left-censored instances. The hyperparameters, the number of chains with different initialization, the number of iterations, length of burn-in period, thinning lag, and the number of central vein segments selected in each iteration were all set to the same values as in Section 4.3.1.

**Performance Evaluation**

Our analysis, presented in Table 4.8, reveals that while the posterior means for parameters are comparable between the right-censoring model and our Gibbs sampler, our method exhibits slightly wider 90% intervals. This broader interval is anticipated as our model accounts for the potential presence of left-censoring. Regarding the recovery of censoring type indicators, both our Gibbs sampler and the right-censoring model, along with an oracle model for comparison, correctly infer the missing censoring type indicators for all censored instances, achieving 100% for both censored and overall accuracy as illustrated in Table 4.7. These results underscore the robustness of our proposed model, demonstrating its ability to perform favorably even when there exists a disparity between the data generation process and model assumptions.

4.3.3 Simulation: Empirical Bias Evaluation

From the results of parameter recovery, summarized in Tables 4.6 and 4.8, we observed that our estimator obtained from the Gibbs sampler significantly underestimates the spatial variance parameter \( \tau^2 \). To investigate whether increasing the sample size would improve the estimate of this parameter, we conducted an empirical bias evaluation.
Table 4.7: Results for predicting the missing censoring type indicator on the data without any left-censoring instances. In the confusion matrix, left-censored, right-censored, and uncensored are abbreviated as "l", "r", and "u" respectively. The overall accuracy encompasses both censored and uncensored cases, whereas censored accuracy specifically considers only censored cases.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Truth</th>
<th>Gibbs Sampler</th>
<th>Right-Cens. Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>0</td>
<td>0</td>
<td>l 0</td>
</tr>
<tr>
<td>r</td>
<td>64</td>
<td>0</td>
<td>r 0</td>
</tr>
<tr>
<td>u</td>
<td>0</td>
<td>238</td>
<td>u 0</td>
</tr>
</tbody>
</table>

Overall Acc. 1 1 1
Cens. Acc. 1 1 1

Table 4.8: The true parameter values of the simulated dataset, which is devoid of left-censored instances. The posterior mean and 90% interval obtained through our proposed Gibbs sampler and through the model only assumes right-censoring.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>Gibbs Mean (90% Interval)</th>
<th>Right Mean (90% Interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>3</td>
<td>2.26 (0.97, 3.67)</td>
<td>2.71 (2, 3.4)</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>-6.4</td>
<td>-6.44 (-6.99, -6.18)</td>
<td>-6.3 (-6.35, -6.26)</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>2</td>
<td>1.5 (0.8, 1.72)</td>
<td>1.68 (1.56, 1.78)</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>10</td>
<td>5.29 (2.96, 5.65)</td>
<td>3.98 (2.64, 5.72)</td>
</tr>
</tbody>
</table>
Simulation Details

We simulated $N$ datasets using the same parameters and spatial frailties, where each dataset consists of $m$ vein segments. The event times for all $N$ datasets are denoted as $\{t_{ij}\forall i \in [m], j \in [N]\}$, with their observed values denoted by $t_{ij}^{[o]}$ (marked as "NA" for the censored instances).

Recall that vein segments approximate xylem vessels, as discussed in "Vein Segment Processing" in Section 4.2.2. Consequently, multiple xylem vessels could correspond to the same vein segment in reality, albeit infrequently. This study can be seen as an initial scenario of this situation, where $N$ xylem vessels are projected to the same vein segment for each segment. Thus, $N$ could be interpreted as the number of events for each vein segment.

Gibbs Sampler for Empirical Bias Evaluation

Under the empirical bias framework, our model is represented in Figure 4.7. When $N = 1$, our model simplifies to the original assumption that at most one event can be observed per vein segment, as shown in Figure 3.2.

The priors are specified as in Section 3.2.1. The likelihood for the $j^{th}$ event on the $i^{th}$ vein segment is:

$$ P(t_{ij}^{[o]} | \beta, \gamma, r_i) = \begin{cases} f_{\beta, \gamma, r_i}(t_{ij}^{[o]}) & \text{if } t_{ij}^{[o]} \neq NA \\ 1 - S_{\beta, \gamma, r_i}(T_1) + S_{\beta, \gamma, r_i}(T_2) & \text{else} \end{cases} $$
The joint distribution of hidden and observed variables now becomes

\[ P(\beta, \gamma, \tau^2, r, t^{[o]}) = \prod_{i=1}^{m} \prod_{j=1}^{N} P(t_{ij}^{[o]} | \beta, \gamma, r_i) P(r-q | \tau^2) P(\tau^2) P(\gamma) P(\beta) \]

Based on the modification in the joint distribution, the full conditional distribution of each variable is adjusted as follows to leverage the information gained from multiple datasets. This enables the effective updating of the related parameters and spatial frailties:

1. \( P(\beta | \tau, \gamma, r, t^{[o]}) \propto \prod_{i=1}^{m} \prod_{j=1}^{N} P(t_{ij}^{[o]} | \beta, \gamma, r_i) P(\beta) \)
2. \( P(\gamma_1 | \beta, \gamma_2, \tau^2, r, t^{[o]}) \propto \prod_{i=1}^{m} \prod_{j=1}^{N} P(t_{ij}^{[o]} | \beta, \gamma, r_i) P(\gamma_1) \)
3. \( P(\gamma_2 | \beta, \gamma_1, \tau^2, r, t^{[o]}) \propto \prod_{i=1}^{m} \prod_{j=1}^{N} P(t_{ij}^{[o]} | \beta, \gamma, r_i) P(\gamma_2) \)
4. \( \tau^{-2(r+1)} \sim \Gamma(a_\tau + \frac{m-1}{2}, b_\tau + \frac{1}{2}(r-q)(D_{(-q)} - a_r E_{(-q)}r_{(-q)})) \), which remains unchanged as compared to the original assumption of \( N = 1 \).
5. \( P(r_{N(i)} | \beta, \gamma, \tau, r_{-N(i)}, t^{[o]}) \propto \prod_{w \in N(i)} \prod_{j \in [N]} P(t_{wj}^{[o]} | \beta, \gamma, r_w) P(t_{qj}^{[o]} | \beta, \gamma, r_q) P(r_{N(i)} | r_{N_{3(i)}}, \tau^2) \)

The hyperparameters, the number of chains with different initialization and thinning lag were all set to the same values as in Section 4.3.1. For our empirical bias study, we used \( N = \{1, 10, 30, 45, 60\} \). To ensure that the number of successful updates for spatial frailties in each iteration during the rejection sampling step remained consistent across different \( N \), we determined the number of central vein segments selected in each iteration through trial and error. Specifically, we set it to \( C = \lceil (0.1 + 0.004(N - 10)) \cdot m \rceil \). Each chain was run for 2000 iterations, with the first 1600 iterations considered as the burn-in period. We ran fewer iterations than those in Section 4.3.1 because, as demonstrated in Figure 4.8, the computational time increases quadratically with larger \( N \).

**Performance Evaluation**

We investigated the behavior of parameter estimates for the spatial variance parameter as the number of events per vein segment increased. Recall that when \( N = 1 \), our Gibbs sampler severely underestimated \( \tau^2 \), resulting in 90% interval not covering the true value (as seen in Table 4.6).
Figure 4.8: The average time per iteration (s) for our Gibbs sampler was evaluated under the empirical bias study across different numbers of events per vein segment ($N$). The dashed line represents the quadratic regression line, with the p-value for the quadratic term being 0.0224, indicating statistical significance. This suggests that the average time increases quadratically with the increase of $N$. 
Figure 4.9: The posterior mean of \( \tau^2 \) obtained from our Gibbs sampler was evaluated across different numbers of events per vein segment (\( N \)) in an empirical bias study. The solid line represents the true value of the spatial variance parameter. Notably, our estimator’s bias decreases as \( N \) increases.

However, with larger values of \( N \), the posterior mean of \( \tau^2 \) converges toward the true value, indicating a reduction in bias. This empirical trend provides evidence of unbiasedness, as demonstrated in Figure 4.9.

We assessed our model’s performance through parameter and censoring type indicator recovery on simulated datasets that mirror the complexities of our application scenarios. Specifically, we evaluated these aspects on datasets both with and without left-censored instances. Additionally, we examined the empirical bias of our model in recovering the spatial variance parameter across simulated datasets with increasing numbers of events. Our findings underscored the model’s accuracy and robustness in parameter estimation, its efficacy in handling temporal gaps, and its ability to yield valuable insights into pivotal events.
4.3.4 Real-world Data

Data Description

We used the same embolism data on leafs as detailed in Section 4.2.2 for evaluating our Gibbs sampler on real-world data.

Implementation Details

In the application to real-world embolism data, we maintained mostly consistency with the simulation settings for the hyperparameter values, the number of selected central vein segments, and the constraint on the number of updated first-order neighbors. Specifically, the hyperparameters are set to the same values as those in Table 4.5, except $\beta_0 = 0$, $B_0 = 45$. We set $C = [0.1 \cdot m]$ and imposed the condition $|N_s(i)| \leq 2$. For each leaf, the Gibbs sampler was executed three times with distinct random initialization. Each chain ran for a total of 5000 iterations, with the initial 4000 iterations considered as the burn-in period. Furthermore, to mitigate autocorrelation and ensure representative samples, a thinning process was implemented, retaining samples at a lag of 40 iterations. This approach aimed to facilitate convergence and yield reliable estimates for subsequent analysis.

Performance Evaluation

We conducted a comprehensive evaluation of our Gibbs sampler on real-world embolism data using multiple approaches. Firstly, we visually validated the correct prediction of the censored midvein as left-censored by our Gibbs sampler. Subsequently, we demonstrated that our data-driven approach yielded results consistent with established phytological findings, such as (a) the observed pattern that thicker veins tend to embolize earlier and (b) the alignment between our fundamental temporal component and the mean $P_{50}$ (in MPa) obtained from optical vulnerability curves. Additionally, our analysis revealed new insights by exploring the relationships between model parameters and vein features, as well as investigating species-level clustering effects for
Figure 4.10: Gibbs sampler result on the *Casearia arborea* leaf showing the relative posterior mean of log odds ratio. The color gradient from red to purple indicates the progression of embolism, where red represents earlier stages and purple represents later stages. Additionally, red indicates left-censoring, while the turquoise to purple spectrum indicates right-censoring.

Visual Validation

While we lack a ground truth for evaluating parameter recovery and censoring type indicator inference, we can visually assess whether censored midveins are now predicted as left-censored. Comparing our Gibbs sampler result in Figure 4.10 with the motivating *Casearia arborea* leaf example in Figure 4.6, we demonstrate the successful prediction of the censored midvein as left-censored. This aligns with phytological findings suggesting that thicker veins tend to experience embolism earlier [20, 21, 22, 23, 24].

Results that align with established phytological findings

The posterior mean of the vein thickness coefficient ($\beta$) from our Gibbs sampler was significantly positive for most cases, with only two exceptions from *Casearia arborea* and *Sloanea berteroana*, indicating that thicker veins tend to embolize earlier (as shown in Figure 4.11). This observation aligns with established literature [20, 21, 22, 23, 24]. It is important to note that our
Figure 4.11: The box plots display parameter values grouped by species.  
(Upper Row and Bottom Left) Species are ordered by Weibull mean, enabling an exploration of the correlation between different parameters and the fundamental temporal component.  
(Bottom Right) Species are ordered by mean vein thickness coefficient, facilitating an inspection of the correlation between the spatial variance parameter and vein thickness coefficient.
finding is entirely data-driven, and we did not impose any prior assumptions regarding the likelihood of the vein thickness coefficient being positive.

The investigation into the species-level correlation between different parameters is depicted in Figure 4.11. In the upper row of Figure 4.11, the vein thickness coefficient ($\beta$) appears to be uncorrelated with the fundamental temporal component, as summarized by the mean of the Weibull family ($e^{-\gamma_1 \Gamma(1+e^{-\gamma_2})}$), and the same holds for the spatial precision ($\tau^{-2}$). The bottom left figure in Figure 4.11 illustrates a positive correlation, as both axes are ordered by the fundamental temporal component. In the bottom right figure, the vein thickness coefficient ($\beta$) appears to be uncorrelated with the spatial precision ($\tau^{-2}$). We also provided investigation on within-species correlation in Appendix A.

In order to validate the theoretical connection established between vulnerability curves and survival functions in Section 4.2.1, we conducted an investigation into the relationship between hydraulic vulnerability, particularly represented by the mean $P_{50}$ (in MPa) as obtained from optical vulnerability curves [35], and various components of embolism propagation in our proposed model. To do this, we calculated between-species Pearson correlation coefficients between a species’ $P_{50}$ and the mean ICAR+th model parameters, which include the mean vein thickness regression coefficient ($\bar{\beta}$), the mean spatial dependence precision parameter ($\bar{\tau}^{-2}$), and the mean of Weibull distribution ($e^{-\gamma_1 \Gamma(1+e^{-\gamma_2})}$). Before computing the between-species Pearson correlation, we conduct the Shapiro-Wilk normality test to ensure that the normality assumptions are met (refer to Appendix B for more details).

In our analysis, it’s crucial to recognize the data structure, particularly the inherent dependence among samples from the same species. The total correlation (i.e., the classic Pearson correlation coefficients computed at individual sample level) is a measure that assumes all samples are independent, neglecting the species-specific structure of the data. This assumption doesn’t hold in our context, as samples within the same species are inherently dependent. Therefore, we must consider either between-species or within-species correlation [40] to accurately reflect the data characteristics.
Figure 4.12: Relationship between mean $P_{50}$ (MPa) from [35] and posterior mean from the Gibbs sampler: (Upper Left) Vein thickness coefficient. (Upper Right) The inverse of the spatial variance parameter. (Bottom Left) The fundamental temporal component characterized by the mean of Weibull family in log scale. The dashed line shows the linear regression line.
Our primary goal is to determine if we can predict a species’ mean parameter value or mean $P_{50}$, rather than solely exploring whether a parameter is correlated with $P_{50}$ for a specific species. As such, we focus on between-species correlation, which takes into account the average parameter values and the mean $P_{50}$ for each species. This approach provides a more comprehensive perspective on the relationships between various components of embolism propagation and hydraulic vulnerability across species, considering the species-level variation and dependencies within species.

Our analysis unveiled a significant between-species correlation between hydraulic vulnerability and the Weibull mean at 90% significance level, while there were no significant between-species correlations observed with other parameters derived from our analyses (Figure 4.12). The observed significant "negative" correlation between hydraulic vulnerability and the Weibull mean is consistent with expectations. A larger Weibull mean implies a tendency for later embolism occurrence, which corresponds to a more negative $P_{50}$ mean. These findings suggest that the fundamental temporal component $S_y(t)$, which encapsulates information akin to the mean $P_{50}$, primarily relates to the temporal dynamics of embolism propagation. In contrast, the spatial and vein thickness components incorporate supplementary information to characterize the spatial and vein thickness aspects of embolism propagation.

**New findings**

In light of the numerous vein features that have been proposed as potential influencers of embolism formation [41], we conducted an investigation into the correlation between vein features and the our model parameters ($\bar{\beta}$, $\bar{\tau}^{-2}$, $\bar{\gamma}_1$, $\bar{\gamma}_2$) at the species level. Specifically, we examined two vein features: average vein density and the degree of vein connectivity. Vein connectivity, a metric frequently used to characterize transportation networks [42], was assessed in our study by quantifying the ratio of branching points to vein segments within the venation network of a leaf.

Our findings, based on our spatial survival model, revealed a significant positive correlation between the vein thickness regression coefficient ($\bar{\beta}$) and the degree of vein connectivity (Figure 4.13) at 90% significance level. However, we did not observe significant correlations between the other model parameters ($\bar{\tau}^{-2}$, $\bar{\gamma}_1$, $\bar{\gamma}_2$) and either of the two vein features under consideration.
Figure 4.13: Relationship between vein thickness regression coefficient from our spatial survival model and (a) Average vein density and (b) Degree of connectivity. The dashed line shows the linear regression results.
In essence, the results suggest that when veins exhibit greater connectivity, the influence of vein thickness on embolism progression becomes more pronounced, as indicated by the larger values of $\bar{\beta}$.

One may inquire whether there is a clustering pattern at the species level for each parameter. This involves examining whether different species have significantly different parameter values, indicating a species-level clustering effect for that parameter. To investigate this, we employed the Kruskal-Wallis rank sum test [43], which is a non-parametric method for one-way ANOVA. The results indicate no significant differences in $\beta$ and $\tau^{-2}$ among the eight species considered. However, significant differences were observed in the temporal parameters ($\gamma_1, \gamma_2$) at a 95% significance level. This implies that only the temporal parameter exhibits a significant variation among different species, aligning with existing phytological findings. While previous phytological work has not explored whether other parameters exhibit species-clustering effects due to the limitations of existing statistical tools, our results suggest the potential for building a hierarchical model with species-level parameters for temporal parameters in the context of embolism formation. See more details in Appendix C.

This real-world application not only confirmed the model’s proficiency in inferring the missing censoring type indicator but also highlighted its practical utility in addressing crucial research inquiries. We demonstrated that the significantly positive vein thickness coefficients align with existing findings that thicker veins tend to embolize earlier. Additionally, among the three main components (temporal, spatial, and vein thickness), only the temporal component exhibited a signifi-
cant correlation with the mean $P_{50}$ obtained from optical vulnerability curves. This result not only validates our model’s ability to capture meaningful temporal information in embolism formation but also indicates its capacity to distinguish between different contributing factors. Furthermore, we discovered that the more interconnected the leaf network, the more pronounced the pattern of thicker veins embolizing earlier becomes. Lastly, species-level clustering effects appeared to exist only for the temporal component, not for the spatial and vein thickness components.

In summary, our Gibbs sampler, evaluated on both simulated and real-world embolism data, consistently demonstrated accuracy and reliability in inferring the missing censoring type indicator. The spatial survival model not only yielded similar findings to existing literature but also uncovered new insights. This contribution provides valuable knowledge, further advancing our understanding of the intricate spatiotemporal dynamics in the embolism domain.
Conclusion

In summary, this thesis is dedicated to the exploration of hydraulic vulnerability using optical vulnerability image data. To engage in spatial survival modeling of embolism events and understand the intricate spatiotemporal dynamics of embolism formation, the first crucial step involves the detection of embolism events from the image data. Hence, we developed a machine learning workflow for embolism event detection. The high sensitivity and accuracy demonstrated by our proposed workflow, assessed on optical vulnerability stem image data, underscore the efficacy of our approach. With the successful detection of embolism events, we proceed to employ spatial survival modeling to unravel the underlying spatiotemporal dynamics shaping embolism formation. The establishment of the equivalence between survival curves and vulnerability curves provides theoretical support for the adoption of survival modeling in this context.

Furthermore, our exhaustive investigation centered on the design and assessment of a spatial survival model, utilizing a Gibbs sampler to tackle the intricacies associated with missing censoring type indicators. The challenge of ascertaining the censoring type for cases marked as censored became particularly pronounced in the study of embolism formation, especially with the advent of optical vulnerability image data obtained through the innovative optical vulnerability technique in hydraulic vulnerability measurement. Our primary objective was to recover the censoring type indicators, thereby providing significant contributions to the nuanced understanding of the intricate spatiotemporal dynamics underlying embolism formation.

The Gibbs sampler exhibited impressive accuracy in the inference of missing censoring type
indicators, showcasing its robust performance on both simulated and real-world embolism datasets. Our evaluation, conducted through parameter and censoring type indicator recovery on the simulated dataset, provided solid confirmation of the model’s efficacy. Compared to spatial survival models assuming only right-censoring, our approach reduced estimation bias and performed well even under scenarios where the data-generating mechanism favors right-censoring. Moreover, the model’s practical utility was underscored by its alignment with established biological findings. This was evident not only in successfully inferring the censored midveins as left-censored but also in the positive vein thickness coefficients, which concurred with the well-documented phenomenon of thicker veins embolizing earlier.

Additionally, our investigation into species-level clustering effects revealed significant differences in the temporal parameters among species, suggesting the potential for hierarchical modeling to enhance our understanding of embolism dynamics. Furthermore, the spatial survival model uncovered a nuanced relationship between leaf network connectivity and the embolism pattern, contributing a novel perspective to the field.

Although our Gibbs sampler was initially developed based on the proportional odds model, its methodology for addressing the issue of missing censorship type indicators can be extended to other survival models. The choice of survival model should be tailored to the specific application. In fact, the post-analysis in Appendix D indicates that the proportional hazard model is adequate for modeling embolism events.

In a nutshell, our spatial survival model, in conjunction with the Gibbs sampler, has proven to be a robust tool for embolism analysis, delivering accurate predictions and uncovering novel insights. Together with the introduced machine learning workflow for event detection, they collectively constitute a unified machine learning framework tailored for the study of embolism events, as depicted in Figure 4.14. This research contributes to advancing our understanding of the intricate interplay between spatial, temporal, and leaf feature factors in the embolism domain, paving the way for more sophisticated and nuanced models in future phytological studies.

A potential avenue for future research involves expanding the machine learning workflow
developed for detecting embolism events in stem image data to encompass leaf image data. The inferred censoring types from our Gibbs sampler could enhance the accuracy of the event detection workflow through transfer learning. Specifically, transferring the knowledge on which vein segments are more likely to embolize earlier could facilitate the event detection phase, leveraging insights gained from the Gibbs sampler to improve performance.

It is crucial to acknowledge a limitation inherent in our approach—the distinction between xylem conduits and vein segments. Our approximation of xylem conduits using vein segments is necessitated by current constraints in image resolution. Future advancements in imaging technology hold the potential to refine this modeling by providing more accurate representations of xylem conduits. Importantly, our spatial survival model coupled with the Gibbs sampler can be easily adapted for xylem conduits when such precise data becomes available.

Another limitation lies in our assumption that each observed embolism event corresponds to a single vein segment. In cases where multiple segments embolize simultaneously within the 2-minute data collection window, particularly in the same regions of the image, we treat them as a singular embolism event. Additionally, we assumed a single vein segment in each censored vein segment. These assumptions may lead to potential underestimation for the actual number of vein segments, especially if overlaying of segments occurs. Although the proportion of observed overlaying events is relatively small, it is conceivable. Future refinements may involve recovering the actual number of vein segments, possibly by considering the association between overlaying
and vein thickness. This avenue presents an opportunity for enhancing the accuracy of our approach in capturing the true complexity of embolism events.

Our study has primarily focused on the application of our Gibbs sampler to embolism data. However, the versatility of our method extends beyond this specific domain. The approach holds potential for application to other time-to-event datasets characterized by an unknown censoring type indicator. Consider, for example, medical scenarios, as discussed in Chapter 1. In such contexts, the utility of a hierarchical model, akin to the one presented in our study, may be even more pronounced. The hierarchical framework could offer valuable insights into the underlying mechanisms governing diverse time-to-event processes, contributing to a broader understanding of complex events across various domains.
References


Appendix A: Within-species Correlation

Examining Figures A.1, A.2, A.3, we observed within-species correlations between the vein thickness coefficient $\beta$ and temporal parameters summarized by the Weibull mean in log scale. On the contrary, the other pairs of parameters did not show within-species correlations.
Figure A.1: Highest posterior density regions for $\beta$ and Weibull mean in log scale.
Figure A.2: Highest posterior density regions for $\tau^2$ and Weibull mean in log scale.
Figure A.3: Highest posterior density regions for $\tau^2$ and $\beta$. 
Appendix B: Shapiro-Wilk Normality Test

The Pearson correlation coefficient assumes that the two variables being correlated follow a roughly normal distribution. Considering the small sample size \( n = 8 \) for the between-species Pearson correlation coefficient, it becomes crucial to verify whether the normality assumption is met. We examined the normality of variables through Shapiro-Wilk normality test. The results, as summarized in Table B.1, reveal p-values all exceeding 0.05. This suggests that the normality assumptions are satisfied for all the variables under consideration.

Table B.1: Shapiro-Wilk normality test on posterior mean of parameters and vein features.

<table>
<thead>
<tr>
<th>Variables</th>
<th>W</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta )</td>
<td>0.8855</td>
<td>0.2125</td>
</tr>
<tr>
<td>( \gamma_1 )</td>
<td>0.9087</td>
<td>0.3448</td>
</tr>
<tr>
<td>( \gamma_2 )</td>
<td>0.9369</td>
<td>0.5812</td>
</tr>
<tr>
<td>( \tau^{-2} )</td>
<td>0.9779</td>
<td>0.9521</td>
</tr>
<tr>
<td>( \log(\text{Weibull mean}) = -\gamma_1 \log(\Gamma(1 + e^{-\gamma_2})) )</td>
<td>0.9033</td>
<td>0.3095</td>
</tr>
<tr>
<td>mean of average vein density</td>
<td>0.9407</td>
<td>0.6182</td>
</tr>
<tr>
<td>mean of degree of connectivity</td>
<td>0.8934</td>
<td>0.2518</td>
</tr>
</tbody>
</table>
Appendix C: Prior and Posterior Distributions

The visual comparison of posterior distributions with prior distributions is presented in Figure C.1. The priors for $\beta, \gamma_1, \gamma_2$ appear weakly informative. However, the posterior for $\tau^2$ exhibits less concentration than the prior, indicating larger spatial variance variation in the data. Additionally, minimal species-level clustering effect is observed on $\beta$ and $\tau^2$, while more pronounced clustering is evident in temporal parameters $\gamma_1$ and $\gamma_2$. 
Figure C.1: The prior distribution, the posterior mean, and the posterior standard deviation for each parameter. The black curve signifies the prior distribution, while the colored circles denote the posterior means. The colored error bars represent one standard error around the posterior means.
Appendix D: Choice of Survival Model for Modeling Embolism Events

While initially designed around the proportional odds model, our Gibbs sampler’s methodology for handling missing censoring type indicators can be adapted to various other survival models. Selecting the appropriate survival model should align closely with the nature of the specific application at hand.

In this study, we compared the proportional odds model with the proportional hazards model to assess their generalization abilities. We used the Deviance Information Criterion (DIC) [44], which evaluates model fitting quality by the negative log likelihood function while adjusting for model complexity. A lower DIC indicates a better model fit. We defined the relative improvement of the proportional hazards model over the proportional odds model as:

\[
\frac{\text{PO model’s DIC} - \text{PH model’s DIC}}{\text{PO model’s DIC}}
\]

A positive relative improvement indicates a better fit of the proportional hazards model.

Figure D.1 demonstrates that the proportional hazards model generally exhibits superior generalization ability compared to the proportional odds model. This suggests that the influence of vein thickness and spatial frailties on embolism events remains relatively consistent over time. This finding points towards a potential future direction for enhancing the generalizability of our Gibbs sampler in modeling embolism events, namely by adopting the proportional hazards model instead of the proportional odds model.
Figure D.1: The box plots display the relative improvement of the proportional hazards model compared to the proportional odds model, grouped by species.